Visual Analytics Methods for Exploring Geographically Networked Phenomena

by

Feng Wang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2017 by the
Graduate Supervisory Committee:

Ross Maciejewski, Chair
Hasan Davulcu
Anthony Grubesic
Paulo Shakarian

ARIZONA STATE UNIVERSITY

May 2017

ABSTRACT

The connections between different entities define different kinds of networks, and many such networked phenomena are influenced by their underlying geographical relationships. By integrating network and geospatial analysis, the goal is to extract information about interaction topologies and the relationships to related geographical constructs. In the recent decades, much work has been done analyzing the dynamics of spatial networks; however, many challenges still remain in this field. First, the development of social media and transportation technologies has greatly reshaped the typologies of communications between different geographical regions. Second, the distance metrics used in spatial analysis should also be enriched with the underlying network information to develop accurate models.

Visual analytics provides methods for data exploration, pattern recognition, and knowledge discovery. However, despite the long history of geovisualizations and network visual analytics, little work has been done to develop visual analytics tools that focus specifically on geographically networked phenomena. This thesis develops a variety of visualization methods to present data values and geospatial network relationships, which enables users to interactively explore the data. Users can investigate the connections in both virtual networks and geospatial networks and the underlying geographical context can be used to improve knowledge discovery. The focus of this thesis is on social media analysis and geographical hotspots optimization. A framework is proposed for social network analysis to unveil the links between social media interactions and their underlying networked geospatial phenomena. This will be combined with a novel hotspot approach to improve hotspot identification and boundary detection with the networks extracted from urban infrastructure. Several real world problems have been analyzed using the proposed visual analytics frameworks. The primary studies and experiments show that visual analytics methods can help analysts explore such data from multiple perspectives and help the knowledge discovery process.

This thesis is dedicated to my family and friends. I couldn't have done this without your support.

First of all, I would like to express my special appreciation and thanks to my advisor Dr. Ross Maciejewski; you have been a invaluable mentor for me. I would also like to thank my committee members, Dr. Hasan Davulcu, Dr. Anthony Grubesic, and Dr. Paulo Shakarian. I would also like to thank all colleagues in the Visual Analytics and Data Exploration Research (VADER) Laboratory at Arizona State University for their support in the past five years, particularly Yafeng Lu and Brett Hansen for your collaborations in my research. Special thanks are also given to Dr. Paul Longley and Dr. Elizabeth Mack, for their cooperation on my two publications. I would also like to thank colleagues in Visual Analytics Team at Pacific Northwest National Laboratory, particularly my mentor, Dr. Aritra Dasgupta, and my manger, Ms. Kristin Cook, for your guidance in my internship. Some of the material presented here was supported by the NSF under Grant No. 1350573 and in part by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001. A special thanks to my family. Words cannot express my gratitude to my mother, father, and grandmother. At the end I would like express appreciation to my beloved wife Mei Hao who was always supporting me even during the hardest days.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Recently, a variety of visual analytics methods have been developed to identify geographic hotspots. These methods typically utilize various forms of aggregation and density estimation to help analysts explore and identify relationships between events and their surroundings. However, such methods often ignore the underlying geographic network features, whether they are social, virtual, or physical network features. Yet, networked phenomena are ubiquitous, everything is connected to everything else [1], and near things tend to be more related than distant things [2]. Thus, by linking network properties to the aggregation model, hidden structures within the data can be revealed. As such, this thesis develops a suite of visual analytics methods for exploring geographically referenced network phenomena.

The major contributions of this thesis are:

1. The design of a visual analytics framework for networked based geographic phenomena;
2. An application of network community aggregation for geographic analysis;
3. The formulation of network based kernel density estimation with variable edge weights, and;
4. A geographic territory extraction methodology from networked phenomena.

These contributions focus on two major approaches: the aggregation of virtual networks into geographic space, and; the aggregation of geographically referenced data onto physical geographic networks. In order to demonstrate the effectiveness of the proposed framework

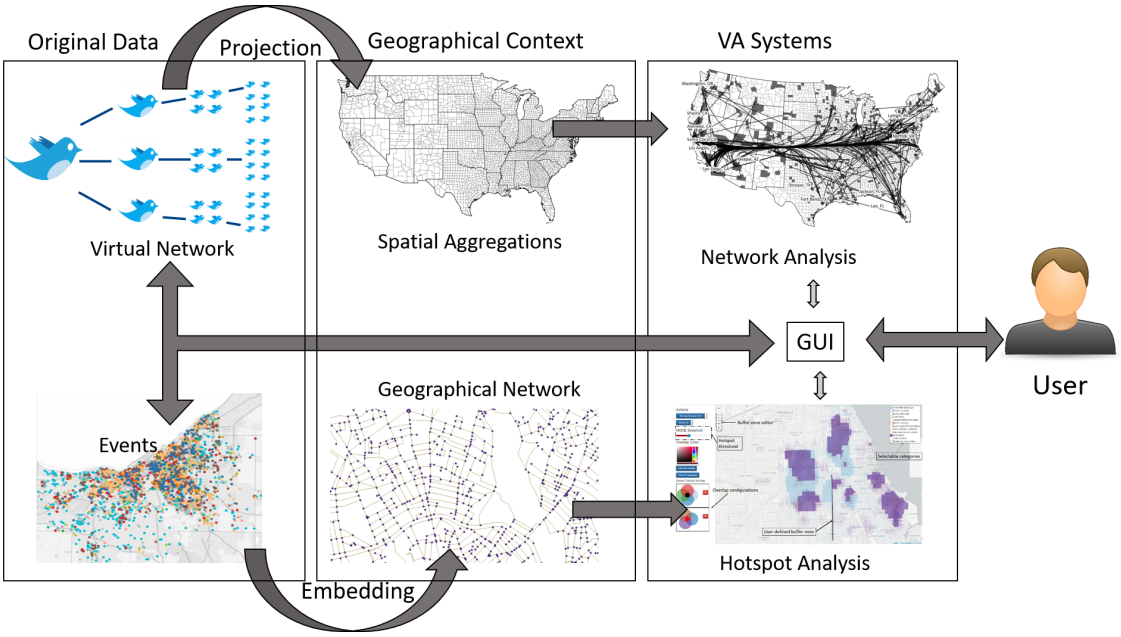


Figure 1.1: Visual Analytics Framework for Networked Geographical Phenomena

and methodologies, this thesis will focus on the application of these techniques to social media data (specifically entrepreneurs) and criminal incident report data.

1.1 Visual Analytics Framework

Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces [3]. Keim et al. explain the pipeline of visual analytics in detail [4]. In the visual analytics feedback loop, the user interactively analyzes the data though a graphical interface (GUI) and automatic models, such as data mining algorithms. This pipeline can be tailored according to the requirements of the analysis. Figure 1.1 provides a conceptual overview of the visual analytics framework used in the proposed system for geographical networked phenomena analysis. First, data are aggregated and linked with their corresponding geographical contexts. This step includes geocoding, record verification, and topology extraction. The geographical contexts are also fed into the visual analytics system for visualizations. The data is then presented through the GUI to the user. Besides the sim-

ple statistics, more information can also be discovered with automatic models. Through the GUI, the user can query the data and change the parameters of visualizations and backend models. The user can also return to the data for verification.

Through this framework, the users are enabled to link the spatial and network aspects in the data. In this thesis, the focus is the visual analytics process provided to the users and how knowledge is discovered through this visual exploration process. This pipeline is applied to social media network analysis under the topic of an entrepreneur network. This case shows how this process can link physical and the virtual networks. This pipeline is also applied to hotspot identification, which is used to demonstrate how our frameworks assist users to link point-phenomena with physical networks in crime analysis.

1.2 Virtual to Physical Network Analysis

One of the aggregation methodologies proposed in this thesis focuses on the aggregation of virtual networks (e.g., online social networks). In recent years, social media has become a test sensor for network expansion, customer relations management, and marketing. User-generated Internet and social media content are forms of big-data which is differentiated from small data by its volume, real time or near real time velocity, and the variety of sources from which it is generated [5]. The first method proposed focuses explicitly on networked phenomenon in virtual networks. The goal here is to aggregate virtual networks and project communities into geographic space to explore spatial associations and patterns. A visual analytics framework has been developed that combines community detection algorithms, geographic projections, and flow maps to enable analysts to link physical and virtual geographies. The goal of this methodology is to support analysts in examining the types of actors engaged in networked phenomena, to examine the geography of locations that underlie these virtual networks, and to analyze the characteristics of these locations as a means of explaining the intensity of virtual phenomena. To demonstrate the proposed

methodology, a case study on social media interactions surrounding entrepreneurship is provided.

1.3 Aggregation Along Physical Networks

Another of the aggregation methodologies proposed in this thesis focuses on the aggregation of data with respect to its underlying physical network geography (e.g., roadways). Here the methodology focuses on how to modify traditional kernel density aggregation models to factor in physical geographic networks. The goal is to aggregate data along physical networks to develop hotspot analysis metrics that better align to an analysts' mental model. A modified kernel density bandwidth is proposed that replaces Euclidean distance with distance traveled along the network. This then allows mapping the density estimation to network properties (such as speed limits), or allowing the analyst to insert a cut in the network based on structural properties; for example, an interstate may serve as a natural boundary to some phenomena. Thus, analysts can extract territories, analyze regions for overlaps between different categories of phenomena, and identify regions with compound risks (e.g., areas where several unique gangs are active). To demonstrate the proposed methodology, several case studies using criminal incident report data are explored.

1.4 Distance and Projections

In geographical information systems, different projections serve for different purposes. All the distance calculations in this thesis are performed using great circle distance, which is the shortest distance between two points on the Earth's sphere. In the local areas, such as the case studies in Chapter 5, this distance can also be estimated with the length of the segment between the projected points. In the larger areas, such as the case studies in Chapter 3 and 4, the distance distortion becomes more obvious so the great circle distance is the only choice. To maintain the distance parallel of latitude, Lambert Conic Conformal projection

is used in the case studies in Chapter 3 and 4. To fit better with the background images and network, WGS 84 Web Mercator projection is used in the local areas investigated in Chapter 5.

Chapter 2

RELATED WORK

The contributions of this thesis focus primarily on network and geographic analysis with a specific focus on the aggregation of virtual networks into geographic space and the aggregation of geographically referenced data onto physical geographic networks. A variety of algorithms for detecting communities and aggregating geographic phenomena have been developed, and the visual analytics community has built frameworks around many of these techniques. In this chapter, related work in social media visual analytics, geographic visualization, network analysis, and hotspot analysis are summarized in order to frame the content of this thesis.

In this thesis, networked geographical relationships will focus on two key aspects: virtual networks bound to projected geography, and point phenomena distribution embedded to physically refereed networks. These frameworks take cues from previous works and are developed to integrate networked phenomena and geographical analysis.

2.1 Visual Analytics for Social Media and Geography

The rise of Web 2.0 and location-based services represents a change in the production of Internet content. These changes have transformed the web from a one-directional phenomenon to a bi-directional collaboration where users are able to interact and provide content [6]. Examples of Web 2.0 services that assemble geographic information include Wikimapia, OpenStreetMap, and GoogleEarth [6]. These new sources of online data are associated with neogeography, which is the use of online data to create maps outside of the boundaries of traditional geographic information systems (GIS) [7]. Several studies have analyzed the content of Internet based geographic applications [8, 9]. Haklay et al.

defines the term “geoweb” as “The Geographic World Wide Web”, which combines the geographical locations and internet topology with techniques such as mash-ups, crowdsourcing, mapping application progamming interface (API), geostack , tags, etc. [10]. The Convoco Foundation visualizes the changes in accessibility of knowledge and the geographical spread of knowledge through the internet in 10 different dimensions [11]. Hollenstein and Purves present a mash-up with images, geo-locations, and meta-data such as tag texts to analyze the tagging behavior and word choices in locations based social network [12]. This study also helps the identification of the fuzzy boundaries between city neighborhoods and areas. This methodolgy is also applied to analyze the relationships between citizen participation in social networks and stereotypes of minorities [13]. Graham and Zoom defines the term “cyberscape” as the representation of geographical locations on the internet and visualize the representations with Google map markers [14]. These studies find an unevenness to the neographies that represent newer manifestations of the digital divide. For example, an analysis of the geography of placemarks and waterlines that were produced to document the damage wrought by Hurricane Katrina reveal persistent socio-economic and racial disparities in the adoption and use of new digital technologies [15].

This case and other studies highlight the power relations behind Internet content which is reinforced by the persistence of large knowledge institutions and the structure of intellectual property rights [16]. These biases raise questions about the representativeness of user-generated data. These representation issues are compounded by the black-box nature of the algorithms used. This nature could happen in the procedure when data are collected and processed before they are downloaded by end users for additional analysis. These issues also create additional algorithmic uncertainties about the procedures, implementations, data models, and structures used to collect and process data [17]. This uncertainty means that spatial patterns generated from big data may in fact be an artifact of the algo-

rithms used to process the data rather than meaningful geographic patterns that reveal new insights about this rapidly evolving world [17].

While social media data is not always a viable way of analyzing all kinds of research questions – particularly those related to unraveling research questions related to interpersonal and social differences [17] – a growing number of studies agree that this new source of data contains promise and has the potential to reveal patterns that are not visible using conventional methods and data [14, 18]. Specifically, social media data can provide resolution on moments and mobility within urban areas at finer time-scales than traditional datasets [19], which facilitates new kinds of analyses about cities [20]. This availability of large digital datasets has been a key factor in the Smart Cities movement [21] where these new sources of digital data are being used to understand cities [21, 22] and provide an exciting entry point for exploring emerging relational geographies. Increasingly, studies of relationships within networks highlight a relational turn in urban and economic geography [23–25]. This relational turn reflects the importance of analyzing the position of cities within different types of networks. Taylor analyzes the world city networks in the hierachical structure framework [26] and three imporant cities which are framed as "command centers" in the network [27]. However, the rapid development of communication and transportation has greatly cut the cost of connections. Zook and Brunn analyze this trend and its influence on the landscape of world city networks by using airline networks database [28]. Although the distance is greatly shortened by the internet, geographical locations still place imporant impacts on cyberspace structures. Tranos and Nijkamp argue that physical distance and other relational proximities still play important roles in internet infrastructure structures [29].

With regard to relational geographies, big data may prove useful in analyzing online social networks. In this respect, the analysis of digitally-mediated interactions on social media is particularly interesting because entrepreneurship is known to be a process where

networking plays a critical role [30–32]. Entrepreneurs rely on networks for a variety of reasons. For example, networks are critical to obtaining access to information, access to customers and suppliers, and financial support [33]. Based on work in sociology which has examined social networks via the strength of interpersonal ties [34], it has been suggested that entrepreneurial networks consist of two primary types of connections or levels: strong and weak ties [32]. Weak ties promote diversity because they are more likely to provide linkages to different groups than would otherwise be connected via strong ties [34]. Thus, weak ties are important because they expand the size of entrepreneurial networks which diversifies access to different types of people and resources [32]. It is also noted that networks need to evolve over time to ensure an optimal number of ties and a diverse mix of actors [35, 36]. This optimal mix is likely to evolve over time between the new venture launch process and subsequent venture development [32]. Overall, studies agree that entrepreneurial networks should be sufficiently large to ensure access to a diversity of information resources. This is important because information is more likely to be spread across a number of individuals rather than concentrated in few resources [37]. Furthermore, recent work on networking shows that social media outlets can help entrepreneurs improve insights about available resources [38], market their businesses [39], and build online social capital [40], although the amount of time spent on these networks can have decreasing returns if utilized too heavily [38].

Despite the wealth of research dealing with entrepreneurship as a networked phenomenon, little is known about the relational geographies between actors within these networks. Physical entrepreneurial networks promote collaboration and the flexibility needed to be competitive in the global economy; it is one of the reasons Silicon Valley is one of the most vibrant locations for entrepreneurship in the world [41]. More recently, however, work has begun to highlight the ability of entrepreneurs to maintain networks cross large physical distances. Saxenians work on the new Argonauts, for example, identifies a group of middle

class people from other countries who are educated in the U.S. and have been successful in stimulating entrepreneurship in newer, more peripheral technology regions through their U.S. based network connections [42]. These networks across large physical distances, and the importance of weak ties in facilitating a wider and more diverse set of actors within social networks [33, 34, 43], provide an interesting backdrop for investigating the relational geographies of entrepreneurial networks.

In recent years, visual analytics has shed light on the analysis of information propagation in online social networks. Whisper is the first comprehensive system to analyze spatiotemporal information diffusion on Twitter by tracing retweeting behaviors [44]. A similiar visual design is also implemented in the system "RApID", which is designed to perform real-time analysis of information diffusion on Twitter with the social graph extraction feature [45]. Based on this direction, more aspects of information spreading patterns are unveiled with data mining and visual analytics approaches. SentiView combines evolution modeling and interactive model adjustments to analyze the temporal dynamics of sentiments in social media discussions [46]. Cody et al. compare the sentiment trends on Twitter indicated by corresponding keywords in a more quantitive approach [47]. Cao et al. propose SocialHelix, which visualizes the sentiment divergence in the discussion about one specific topic on social media [48]. By using data mining algorithms for outlier detection, Zhao et al. propose a visual exploration framework, FluxFlow, which shows the anomalous information spreading [49]. Xu et al. visualize the competitions between topics on social media [50]. As a followup, Sun et al. propose EvoRiver, which is a visual analytics system combining cooperation and competition relations between topics in a modified theme river visualization [51]. Wu et al. propose OpinionFlow, which visually summarizes diffusion of opinions about topics by integrating information diffusion models and density maps [52].

The GPS coordinates embedded in the social media posts can also be used as a sensor for events. By aggregating the locations in the social media posts, we can discover

movement patterns in urban commuting networks [53, 54]. ScatterBlogs2 provides a visual exploration tool to perform real-time monitoring over social media text, geolocations, and topics [55]. SensePlace2 is proposed to show the potential in closing the gaps between geographical analytics on Twitter, event detection, and crisis management [56]. Abnormal events are detected through real-time monitoring of such sparse information [57]. By extracting the trajectories of the specified users, anomalous movement can also be detected [58]. Thom et al. propose a mash-up system to show the related social media keywords and their geographical hotspots to detect anomalies [59]. LeadLine provides linked views to help event detection though temporal bursts, keywords, and geo-locations [60]. Social media data can also effectively track spatially spreading events, such as earthquakes [44, 61].

2.2 Network Based Geographic Visualization

There are two major categories of visualizations for network based geographic visualization. The first category derives from graph drawing algorithms which use node-link or adjacency matrices. The second category starts from networks, which are physically constrained in their geographical background, and embeds information into the networks.

The most straightforward visualization for graphs is a node-link visualization. This visualization is intuitive, straightforward for global structure with proper layouts, and very flexible. However, it has high complexity (greater than N^2) for layouts and is unscalable to large networks because of computational complexity and visual clutter. Solutions to these issues include clustering, bundling, crossing reduction, and filtering. Clustering and bundling methods reduce visual clutter by combining nodes or edges together. Filtering methods take the network structure to guide the user on subnetworks to simplify the analysis. Filters often rely on clustering results. Another challenge of node-link visualizations is binding nodes to geographical locations. The first solution is to use 3D visualizations to

avoid the positional restrictions of the 2D geographical space [62]. The added difficulty of visual tracing and interactions do not yield obvious advantages in traditional 3D displays. To solve this issue, new interaction techniques such as Virtual Reality (VR) can be introduced [63]. Another method is edge bundling. Edge bundling can be based on metrics such as geometry features [64, 65], data hierarchies [66], force directed interactions [67], and kernel density estimation [68]. When the graph is not very large, proper arrangement of edges can also reduce clutter introduced from intersections and overlaps. Cartography surveys provide design suggestions for graph drawing in geographic visualizations [69, 70]. They suggest that the visualizations should be adapted according to the priorities in analytics tasks. For example, in underground railway maps, the topology of connections should be prioritized by relaxing the physical location restrictions. Another suggestion is to use multiple linked views.

An adjacency matrix is another visualization method for networks. This visualization is suited better for dense graphs because it avoids edge crossings in node-link drawings. However, it is not as intuitive as node-link visualizations. It is also very hard to track a path in a matrix visualization. This visualization is more suitable to track the quantitative edge properties in a network. A typical application of an adjacency matrix visualization in geographical networked data visualization is origin-destination (OD) matrix visualization [71]. By putting OD matrices onto the map, the OD map shows the associations between geolocations and travel patterns [72]. Maptrix combines the OD matrix and the linked map view to show many-to-many flows among geographical units [73].

Another trending geographical networked data visualization is to embed information into physical networks. These visualizations are required mostly in trajectory visualizations. Andrienko et al. present a survey of possible aggregation methods of movement data [74]. There are also two major types of the integration between spatial movements and other dimensions. The first method uses 3D visualization to add a dimension besides

geographical locations. This idea comes from the concept of a 3D spatio-temporal cube so it is intuitive and good for presentations. However, it also increases lots of complexities to visual observation on local areas and interactions. An example is the stacking-based visualizations of trajectory data proposed by Tominski et al. [75]. For example, temporal values can be "stacked" into wall-like visualizations. Another kind of 3D integration is inspired from vessel visualization in medical imaging which renders the network as vessel-like shapes [76]. This method can also be extended with density functions [77]. The second method uses 2D visualizations and encodes the information with shapes or colors on the roads. To embed shapes, such as charts, into road maps, deformation is necessary because of the limited of visual road width. Along a specific route, Sun et al. propose two algorithms to widen the roads to embed temporal theme river [78, 79]. This visualization suits detailed temporal information along a specific route. However, it cannot present the global spatial patterns. With the density function, Bristle maps encode values into shapes along the roads, which enables multivariable density visualization [80].

Besides the visualizations summarized above, interactions can also assist the analysis procedure. For example, a fish eye lens can dynamically zoom into the selected area for more details. Edgelens utilizes an algorithm which pushes the edges in the fish eye lens away from each other to make them more visible [81]. TrajectoryLenses aggregates trajectories and uses an interactive lens for filtering and detailed information [82]. Liu et al. visualize temporal information with a circular axis around the selected area [83]. Hu et al. propose a two-layer 3D visualization to integrate a force-directed graph layout and geographical context [84]. Linked multiview is also helpful to understand the complex attributes along spatial networks [85].

2.3 Network Analysis Algorithms

To analyze the structure of the virtual networks and their relations to see underlying geographical context, one fundamental approach is to identify clusters or communities in the networks. In most cases, the communities are identified with regard to an objective function. This function defines the intuition of a cluster. In most analytic tasks, the clusters are defined as groups of nodes with better internal connectivity than external connectivity. Leskovec et al. compare a range of objective functions [86]. Girvan and Newman propose an algorithm based on "edge betweenness", the GN algorithm [87]. The GN algorithm avoids many shortcomings and was used as a standard method in network community identification. However, this algorithm is very expensive in terms of compuational complexity. The complexity is M^2N, where M is the number of edges and N is the number of nodes. This complexity makes it pratically infeasible for large networks [88]. Using the quantitative definition — which is similiar to modularity — Radicchi et al. propose an algorithm using structural edge-clustering coefficient based on triads, which has lower computational complexity than the GN algorithm [89]. Modularity is another straightforward metric of the quality of an partition [88]. It compares the density of inter and intra-community edges. However, it is NP-hard to reach the global optimization of the objective function [90–92].

From physics, the ideas of conductance is introduced into social network analysis. Kannan et al. propose a spectral algorithm which maximizes the connectivity and resistance of the inter-community edges [93]. Flake et al. propose a heuristic based on network flow construction [94]. They also define the community in a network as a set of nodes which have more edges pointing to the inside of the community than outside [95]. This pair-counting idea is applied in defining strong and weak communities [89]. This idea is also used as a comparison metric for different clustering results [96]. As a followup, modularity is introduced as a measure for an optimization based network community detection [97].

Besides the hierarchical optimization procedure, spectral approaches are also used to get the approximate results [98, 99]. This thesis focuses on the algorithms based on hierarchical modularity optimization. Xiong et al. propose an information diffusion model and discovered that the modularity in social network topology impacts the outbreak size and spreading speed [100].

2.4 Geographic Hotspot Analysis

Besides projecting a virual network onto its geographical context, the second aspect of geographically networked data analysis is to analyze event distribution patterns bound to physical networks. Hotspots, defined as areas of concentrated incidents, are widely used for spatial analysis and predictions (e.g. [101–103]). The most widely implemented method is Kernel Density Estimation (KDE) [104]. Implementations of KDE are available in a variety of tools which have been used to study wildfires [105], air traffic patterns [106], criminal incident reports [107], etc. For example, Maciejewski et al. explore spatiotemporal changes in emergency department records using a kernel density estimation [108]. Scheepens et al. apply kernel density estimation to trajectory aggregation and use contour lines to help analysts predict the movement of ships [76, 109]. Tao et al. transform hotspot into hot flow detection with kernel density functions over flow data [110]. Razip el al. present a mobile toolkit to help citizen and law enforcers assess risk levels in urban areas where risk was visualized as a density estimation [111]. Work by Krisp and Peters focuses on using density estimation for vectors to reflect temporal movement trends [112], and Kim et al. extend kernel density estimate heatmaps by introducing linear elements in the map [80]. According to a report by the National Institute of Justice, hotspots can be identified from various methods and they should be chosen according to the tasks of the applications [102]. The identification of spatiotemporal distribution patterns can provide guidance to interventions for crime reduction [113].

As such, previous studies have demonstrated that the underlying assumption of a 2D space is too strong when exploring point events occurring in a geographically defined network space [114, 115]. Given these issues of spatial boundaries and Euclidean distance measures, there is a need for techniques that can better incorporate geographic features, particularly networks. This has led to a variety of methods that focus on incorporating the underlying network geography. Okabe et al. applies kernel density functions to geographical networks, which were later implemented in SANET, a toolkit for geographical network analysis [116–118]. Xie and Yan proposed lixel, a discretization approach for network KDE, and made a series of experiments to compare the effects of different kernel functions and parameters [115]. Borruso explores the effectiveness of network KDE in real world applications [119], and Shiode introduces geographic network into spatial clustering [120]. Tompson et al. proposed hot routes to investigate the density of crime along daily commute routines in London [121], and Heim developed a segment-indexing approach to analyze street crimes [122]. Laffan and Taylor introduce boundaries into the KDE calculation based on user drawn routes [123]. BirdVis integrates tag clouds and KDE heatmap for the analysis of relative habitat preferences [124].

Specifically, hotspots are often used in crime prevention. The core hypothesis is that crimes are concentrated in certain places [125]. In practice, crimes are concentrated in certain street segments. Andersen et al. find that 50% of property crimes happen on only 5% of street segments and intersections in Vancouver [126]. However, 5% of street segments cannot provide direct guidance in large metro areas because there are still too many street segments. Hipp and Kim propose a solution which uses nonuniform spread analysis of crime incident data [127]. Hotspots are successfully applied in police patrol scheduling [128–131] and public policy making [132, 133].

There are several concerns about the accuracy and effectiveness of KDE hotspot methods in spatial analysis. The first concern is the data quality. Hart and Zandbergen show that

hotspot mapping can greatly alleviate the data quality issues introduced by geocoding errors and GPS device errors [134]. Chainey et al. propose prediction accuracy index (PAI) for hotspot identification methods to perform the quantitative evaluations [135]. Drawve et al. compare different hotspot mapping and modeling methods with PAI and Recapture Rate Index (RRI). They conclude that integrating the environmental background can improve overall performance [136, 137]. Based on the evaluations and user studies, Ratcliffe proposes a framework to guide better crime prevention strategies based on the spatial hotspot properties and underlying temporal trend patterns [113]. The hotspot algorithm proposed in this thesis is also evaluated under different settings.

Chapter 3

VIRTUAL NETWORKS IN GEOGRAPHIC SPACE

Geographical space influences virtual spaces such as social networks with spatial distance [1]. A visual analytics framework is proposed to analyze the interactions between different geographical areas in virtual space. Social networks constructed from US entrepreneurs are used as a demonstration.

These data are promising because they represent untapped sources of information which may help us understand the rapidly evolving world, and a growing body of work has highlighted their analytical utility [11, 18, 138] due to the high spatio-temporal resolution they provide [19]. This thesis focuses on networked phenomenon in virtual networks and their projections into geospatial space. Specifically, this work focuses on an the example of social media interactions surrounding entrepreneurships.

Despite the promise of social media data, the sheer volume of information generated present storage and analytical challenges [139] as well as challenges in quality control, documentation of content, documentation of data collection techniques, and rigorous sampling methodologies [5, 140]. These issues call into question the representativeness of analytical results drawn from data sources [17, 141, 142] and raise important questions about the appropriate research domains and conclusions that may be drawn from the data. However, with understanding process of who uses particular social media platforms, this user information represents a way forward for understanding the types of research questions for which social media data may prove valuable [143]. For example, social media user profiles that match with populations of research interest may be a way of mitigating (albeit not completely eliminating) representation issues associated with social network data. In this regard, the use of social media data in order to analyze entrepreneurial networks represents

a promising research area where the data from Twitter could prove valuable. Prior work has highlighted that the process of conceiving of and starting an entrepreneurial venture is a fundamentally networked activity [30–32] where weak ties from a diverse and large group of actors can improve access to information [37]. More recent work also highlights that social media can improve access to information for entrepreneurs [38].

Given the promise social media offers for entrepreneurship studies, this thesis explains methods to analyze digitally mediated interactions using Twitter data collected about a variety of actors engaged in entrepreneurial networks for the United States over an eighteen-month period. The goal of this analysis is threefold: to examine the types of actors engaged in these digital networks, to examine the geography of locales active on Twitter in the entrepreneurship domain, and to analyze the characteristics of locales that explain the intensity of activity on this social media platform. This study makes two important contributions to entrepreneurship research and geographic research on social media. First, the analysis highlights how social media may be used to analyze entrepreneurial social networks on a larger scale than social network data generated from primary survey data collection efforts, which has limited prior work on these networks to case studies [31, 144]. Second, the incorporation of ancillary data in the analysis highlights how social media may be used to explore relational geographies of entrepreneurship and the characteristics of places interacting via social media and other types of digitally-mediated communication platforms [143]. Analytical results reveal that the hashtags used in this analysis do capture (albeit not exhaustively) well-known, important actors in entrepreneurial networks and important subtleties in the geography of locales engaged in these networks. The network clustering algorithm used in the analysis is based on modularity optimization. In the next chapter, the effectiveness of different algorithms are also compared based on modularity optimization on this dataset.

3.1 Data Description

The United States was selected for analysis because it is regarded as a country with numerous hubs of entrepreneurial activity including Silicon Valley in California, the Route 128 area of Boston, Massachusetts [41, 145], and Austin, Texas. For the purposes of this study, entrepreneurial activity is characterized by retweets that are identified as pertaining to entrepreneurship by the hashtags #entrepreneur or #smallbiz [1] . These hashtags were selected because they appear frequently in Tweets from people that identify as entrepreneurs as well as Twitter communications from entrepreneurship support organizations such as the Kauffman Foundation and the Small Business Administration. In addition to the work done by the authors, several online articles that talk about Twitter and entrepreneurship explicitly highlight that the hashtags used in this paper are important for entrepreneurs to follow and use [146–148]. Another advantage of hashtags over plain words is that hashtags can exclude many unrelated tweets which can be confused with similar word choices. For example, the keyword ”small business” might include results such as ”small business card printing”. The major analysis in this study focuses on retweets based on the precedent of prior work [143]. Retweets are an indication of the spread of information within networks and have been used to analyze the influence of particular users within networks [149]. These ideas are integrated in a geosocial gauge, which is a system to summarize information in geo-located tweets [150].

This study evaluates networked geographies of entrepreneurial activity between U.S. counties. Information gleaned from retweets are designed to capture interactions between formal and informal actors in entrepreneurial networks, per Birley who defines entrepreneurial networks as containing “formal networks (banks, accountants, lawyers, Small Business Ad-

[1] This thesis argues that small businesses are distinct from new entrepreneurial ventures and the use of the hashtag #smallbiz in combination with the hashtag #entrepreneur is not meant to suggest otherwise. Instead, the combination of these hashtags comes from their frequent use in combination with one another and also the fact that most people (even entrepreneurs) confuse these two types of enterprises.

Twitter Screen Name	Name
Support Organizations	
azcommerce	AZ Commerce Authority
CommerceGov	U.S. Commerce Dept.
HouseCommerce	Energy and Commerce
USDOCLibrary	Commerce Library
HBCUChamber	HBCU Chamber of Commerce
AustinEcommerce	eCommerce Websites
USDOL	U.S. Labor Department
KauffmanFDN	Kauffman Foundation
1MillionCupsPRO	1 Million Cups Provo
Information Resources	
Inc	Inc.
WSJsmallbiz	WSJ Small Business
JOC_Updates	Journal of Commerce
FastCompany	Fast Company
FastTrac	Kauffman FastTrac
EntMagazine	Entrepreneur
Venture Capital/Financial Resources	
UMSocialVenture	University of Michigan Social Venture Fund
UrbancapitalChi	Urban Capital Chicago
kelvincapital	Kelvin Capital
ingresscapital	Ingress Capital
licapital	Long Island Capital Alliance
kpcb	Kleiner Perkins

Table 3.1: Entities Active on Twitter

Figure 3.1: Sample Tweets in the Topics of Entrepreneurship Geolocated Using User Profiles

ministration (SBA)) and the informal networks (family, friends, business contacts)" [31]. The retweets collected are also designed to capture entrepreneurial activity across multiple stages of the venture creation process (motivation phase, planning phase, and establishment phase) [151]. Table 3.1 contains a sample of different types of entities active on Twitter for this particular study period and validates that the hashtags used in the ensuing analysis do indeed capture interactions amongst entrepreneurial actors across multiple phases of the venture creation process. This table contains the names of support organizations, information resources for entrepreneurs, as well as venture capital companies and other sources of financing for new ventures. Individual entrepreneurs, as well as other persons active on Twitter and engaged in the entrepreneurship community, are not listed in this table due to privacy concerns. Inc. magazine, for example, is a magazine focused on issues pertaining to entrepreneurship and small businesses. It is also famous for its annual publication of Inc. 5000 firms, which is a list of the fastest growing small, private U.S. businesses [152].

Figure 3.1 displays some sample tweets collected for this study. In this figure, the user name and profile photo are in the top-left corner, and specific profile details (user specified location, followers) are obtained by clicking the user name. Tweets also contain specific grammar including hashtags, mentions and retweets. Hashtags are words or phrases begin-

ning with the "#" sign. This is used to mark topics associated with the tweet. Mentions use the "@" symbol with a user name to link the Tweet to another user. Retweets are a reposting of someone elses Tweet in a Twitter feed with the "RT" symbol as an attribution. For this study, Twitter data with the hashtags #smallbiz and #entrepreneur between January 1, 2013 and June 30, 2014 were collected. If a Tweet is retweeted, all retweeted incidents are also captured from the stream as the retweet will contain the same hashtags of interest (i.e., #smallbiz and #entrepreneur).

To collect tweets, two crawlers were developed to combine the Topsy and Twitter search APIs. Here, the use of the Topsy API is critical because this company is a certified reseller of Twitter information that provides full access to the Twitter stream, which is an improvement over the one percent sample of tweets provided by the Twitter streaming API [153]. First, the crawler queries Topsy for all Tweets within a given time range, and the results from Topsy are returned. Next, the crawler results are fed into a distributed Twitter crawler based on the Twitter Search API [154]. The Twitter crawler searches the tweets by the ID number from Topsy results and saves the full Tweet into a MongoDB instance. Based on this process, 6,507,506 tweets were collected. 24.5 percent of these tweets (1,594,530) are retweets that contained the hashtags #smallbiz and #entrepreneur. These 1,594,530 tweets are used as the basis for creating the retweet network which represent interactions on Twitter.

3.2 Geocoding Tweets

The tweets collected are linked to 297,639 unique users. To geocode tweets, locations are derived from profiles based on either GPS coordinates or text information. This is possible because users share locations in their profile as free-form text, such as "NYC" or "Bay Area, San Francisco" which can be converted into coordinates with map APIs as is the case with prior studies [143]. User profile information is the principle means of location

information in this study because we are interested in associating tweets with the locations users identify as their base of operations. Although users can lie about their locations or provide unusable information such as "universe" in profile information [16, 155], we argue that active users on Twitter engaged in entrepreneurship have an added incentive to provide accurate location information; particularly if they are using social media to market their businesses as prior studies have indicated [39].

Users and their associated tweets are geocoded to the county level per the precedent of prior work [142] and to minimize geocoding errors that could propagate at finer scales of analysis. Of the users in the global dataset, 89, 914 were ultimately geocoded to county level locations in the U.S. The vast majority of these users (89, 411) were located from user profile information, the remaining users (503) were located from GPS coordinates. A small fraction of users (less than 0.5 percent) had multiple cities listed in their profiles (e.g. Milwaukee/Madison/Chicago). In these instances the first city listed was used to geocode the user. Given the noted mismatch between GPS coordinates and user profile information, which may indicate the location of the person tweeting instead of the GPS location of their base location [16, 143], the match between GPS locations and user profile information was also examined; there was an 82% match between users with both GPS and location information in their profiles Overall, 5.5 percent (126,366) of all tweets collected could be geocoded. While small, this percentage is five times the one percent of tweets typically used in studies where only GPS identified Tweets are used for geographical information. A summary of the users and tweets can be found in Table 3.2 where shaded cells sum to the total number of users in the dataset [153, 156].

3.3 Representation Issues with Twitter Data

As with any data source, there are questions associated with the representativeness and accuracy of these data. This is particularly true with Twitter data, which has been noted to

represent a very specific sub-set of people [141, 142], which limits the utility of these data for the analysis of social, gender, and racial issues [17]. Other issues noted with these data are that user accounts are not unique to one user but may be used by multiple people, or one person may have multiple accounts [141]. Another issue noted with using Twitter data is that it may exclude users that are active listeners and participants in the network but do not actively post information on the network [157, 158]. Studies do suggest, however, that Twitter data may be used to understand Internet mediated interactions in a meaningful way when combined with ancillary quantitative and qualitative data [143]. Twitter and internet users are widely used as sensors for social events and opinions. However, Twitter data have sampling biases. First, Twitter users are more likely to be from populous areas. People in sparsely populated areas are more likely to be underrepresented. The second bias is the gender bias declined to male users [142].

In an effort to mitigate representation issues associated with the small sample of Twitter data that may be geocoded, the use of data from Topsy, which provides access to the full Twitter firehose, mitigates well-known sampling issues associated with use of data from the Twitter API [153]. Further analyses on the data were also performed to check for potential pattern bias due to a large volume of Tweets generated from particularly dominant users. An analysis of Tweets per user revealed an average number of 1.4 tweets per user and the distribution follows the known power law distribution of Tweets. Users with a large number of Tweets [2] in the dataset were further verified to ensure they were reputable sources. A harder issue to grapple with, however, is the issue of social, economic, and demographic representativeness of these data. It is well-established, for example, that user-generated data is dominated by highly educated, Western, wealthy, white males [143]. Despite these issues, demographic studies of the Twitter user community suggest that, in

[2]The user most retweeted by U.S. users is @AlleyWatch. The user with the most retweets from U.S. users is @InnovateM.

the context of applying Twitter data to studying interactions between members of the entrepreneurship community, these issues of representation are less problematic than other applications of Twitter-derived data because the profile of Twitter users and entrepreneurs overlaps substantially.

According to a recent Pew Research survey, twenty-three percent of online adults used Twitter in their March-April, 2015 survey [159]. This report highlights that thirty percent of the users were urban residents, twenty-one percent suburban, and fifteen percent rural. It also highlights that the majority of users are higher income adults under the age of 50. Outside of this study, other work evaluating the demographics of Twitter users has found supplemental and supporting evidence to the Pew study. Adnan et al. found an underrepresentation of ethnic minorities on Twitter, particularly Hispanic/Latinos [160]. They also found that the largest segment of the Twitter user base is in the 20-40 age bracket. Aside from these demographic characteristics, other work has found unusually high representations of creative occupations (actors, artists, and writers) who use Twitter for promotional purposes [161].

Interestingly, the demographic of Twitter users overlaps substantially with the profile of entrepreneurs. It is well established that while older people do participate in entrepreneurship, their willingness to do so declines with age [162, 163]. Thus, younger people are more likely to be entrepreneurs than are older people [164, 165]. In fact, studies suggest that the prime age for entrepreneurship is between the ages of 30-40 [166]. Based on these studies and others, which find that entrepreneurs are white, highly educated males between the ages of 35 and 64 [167], there does appear to be a substantial association between the profile of entrepreneurs and the Twitter users. Entrepreneurs are also noted to use Twitter for business networking purposes [38, 168].

3.4 Assessment of Topic Discussions in Hashtag Cohort

Along with the issue of representativeness, it is also important to assess that the discussions on Twitter around the chosen hashtags are relevant to entrepreneurship. Using all tweets in the dataset from the 89,914 users geocoded as previously described, topic modeling analysis was conducted using latent dirichlet allocation (LDA) [169]. LDA takes each Tweet and attempts to learn the top k topics based on the probability of word occurrence in the Tweets. A value of k = 20 was used to extract the 20 largest topics from the corpus of tweets. The largest topic was motivational/inspirational leadership quotes, followed by tips on building a business website and blog, discussion on hot brands in the market, and tips for small business owners. The only major topic discovered that was marginally unrelated was a topic relating free giveaways; however, these giveaways can be seen as marketing strategies for small businesses to get their products to market. Overall, this analysis indicates that the topics found in the retweets related to #entrepreneur seem relevant to discussions of entrepreneurship and would fall into the concept of capturing entrepreneurial activity across multiple stages of the venture creation process (motivation phase, planning phase, and establishment phase) [151], as discussed previously.

3.5 Method

In order to explore the networked interactions between Twitter users engaged in entrepreneurial networks, a suite of analytical techniques are combined including network analytics, community detection, and regression analysis. This methodology expands upon previous work with TweetXplorer [170], which was used for exploring either the number of tweets, or the social media network, but not both simultaneously. In this analysis, ancillary county data are also incorporated to contextualize the network analysis as recommended by [143].

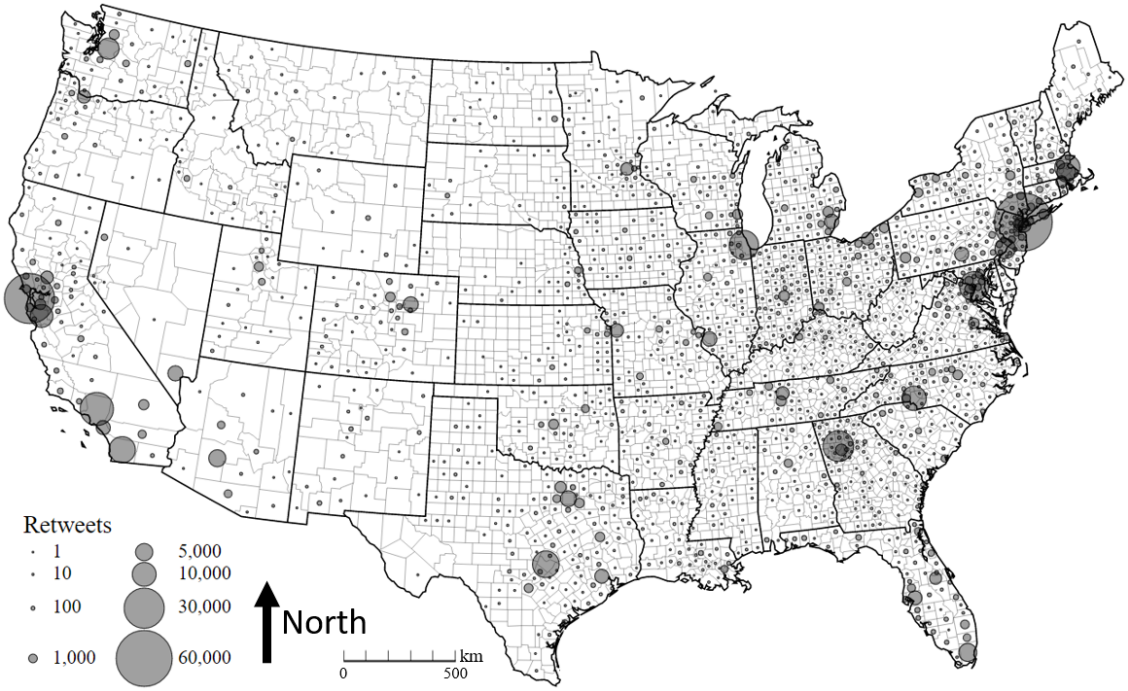


Figure 3.2: Total Retweets per County

3.5.1 Network Analysis

In the network analysis, the retweets between users are considered to be directed edges between the counties. In this way, it is possible to explore the networked geographies of users. If a Tweet originates in county A and is retweeted by county B, then the direction is considered to be A flowing to B. Edges flowing in the same direction between two counties are merged into one weighted edge, the weight of which is the number of the retweets. Once the graph is created from the data, the Clauset-Newman-Moore algorithm [171] is applied to identify network communities in the data. This is a hierarchical aggregation algorithm for detecting community structures within networks using a greedy optimization based on the modularity of a network. Modularity Q is a measure of the strength of connectedness within a network, and social connections in communities imply a faster rate of transmission of information. As such, communities in networks are defined as groups (or clusters) of

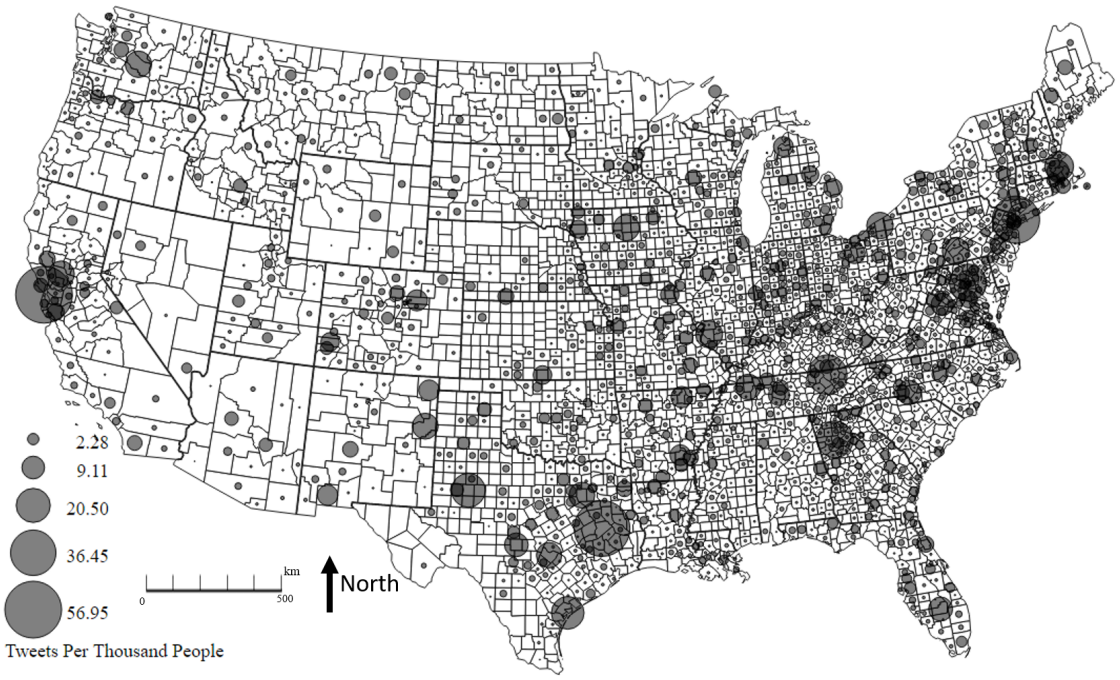


Figure 3.3: Retweets per Thousands of People for Each County

densely interconnected nodes that are only sparsely connected with the rest of the network. The definitions of modularity are explained in detail in the next chapter. High modularity values correspond to well-defined community structures.

The Clauset-Newman-Moore algorithm begins with each node in the network representing its own community of size one. The algorithm then repeatedly joins together two communities such that the resulting new community will represent the largest increase in the modularity value, Q. Thus, for a network of n users, $n-1$ combinations will leave us with a single community. The algorithm converges when the modularity cannot increase anymore or it reaches a user-defined threshold. The threshold can be selected by cross testing on different parts of the data. First, the data is split into segments and the algorithm computed without using a threshold. In this case, the algorithm will converge into one community. Then sub-communities are discovered by reversing the computation pro-

cedure. In this manner, small communities may be identified, as well as other larger major communities. By observation, the maximum polarity becomes stable near a threshold value of 0.2. In experiments on the dataset, the smallest major community consists of 98 counties and the largest minor community (noise) consists of 4 counties. Based on this analysis, a threshold value of 0.2 is used to extract communities.

This process can be compared to a random graph which has no community structure. From the definition of modularity, a random graph has a modularity of 0. Any deviations from the random graph will lead to a clustering effect that identifies communities, which enlarge the value of Q. As Q increases during the algorithms procedure, clusters form, which represent the community structures of the graph. In other words, the clusters found in this process represent network community structures. This means that counties within the same cluster are more likely to have stronger internal interactions than counties in different clusters. Furthermore, the connections between the clusters will be sparse, and by filtering the edges based on centrality measures, analysts can quickly identify the central hubs of the clusters. Once a cluster classification is assigned to a county, the networks can be projected onto a choropleth map [67] as shown in Figures 3.4−3.9.

3.5.2 Visual Analysis of Network Interactions

In order to explore relational geographies of Twitter communications related to entrepreneurial activity, a visual analytics (VA) framework combining geographic community projection, and flow maps was developed. Flow maps combine maps and flow charts as a means of showing the movement of objects from one location to another, making them ideal for communicating retweet flows from the region in which the Tweet was generated to the regions in which the retweet was generated. The VA framework in this paper combines network community detection using the Clauset-Newman Method described in Section 3.5.1 and flow map visualizations.

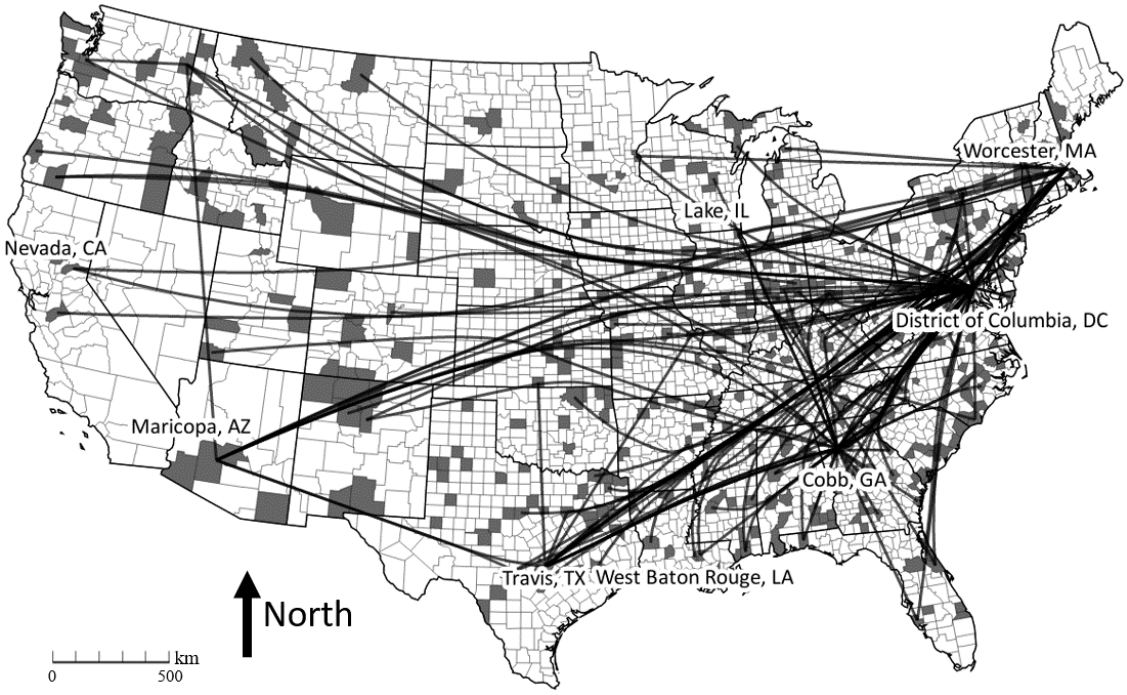


Figure 3.4: Community 1 Includes the Key Nodes Including Suffolk, MA, Fairfax, VA, and Cobb, GA.

The flow maps are visualized as node-link diagrams in geographical space. However, node-link diagrams with a large amount of nodes and edges often suffer from visual clutter. For the balance between interactive performance and visual effect, a force-directed edge bundling for flow map simplification is used. This algorithm models the edges as springs that can attract each other. The edges are divided into subdivision points. Within one edge, zero-length springs exist between consecutive points. Between nearby edges, an attracting electrostatic force is used to combine edges. In each step, the points are moved according to the forces calculated from the two forces. Users can also interactively filter edges based on network properties such as in-degree, out-degree, and centrality. In this manner, analysts can quickly explore both internal structures of the communities as well as filter to see the most salient community hubs, as shown in Figures 3.4-3.9. To find the spatial relationships

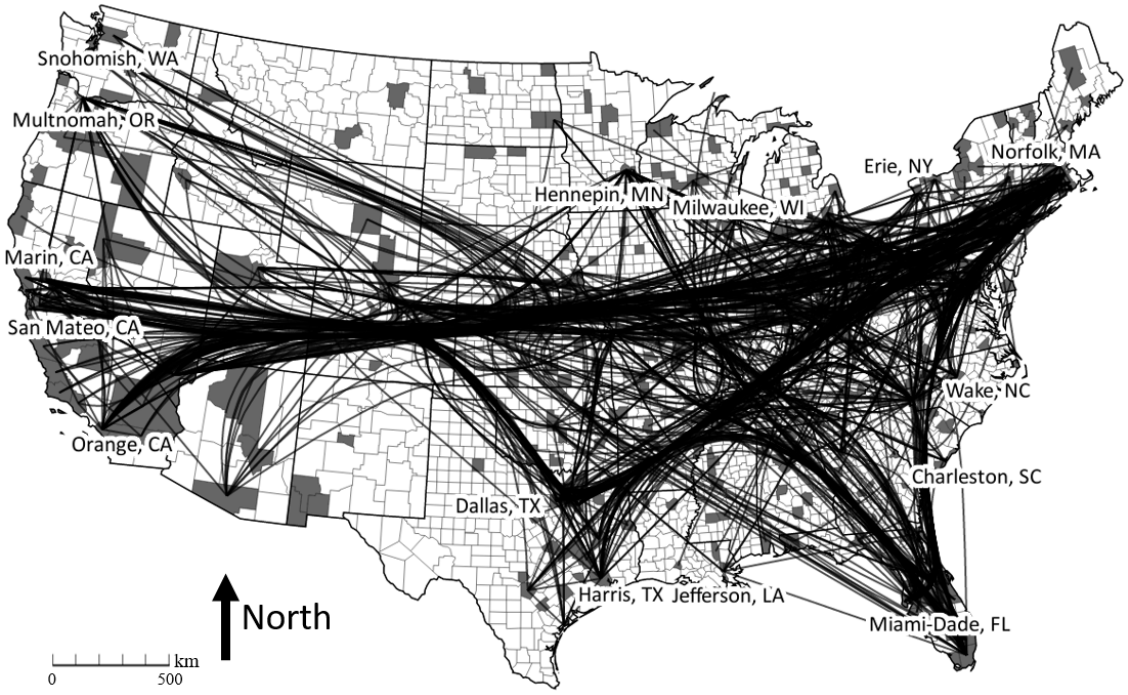


Figure 3.5: Community 2 Covers the Locales on the East and West Coasts, in the West, and in the Midwest.

among counties within a specific community, the average distance from the county to other counties in its community is calculated. The links within the communities are also drawn on the map to show the spatial structure of the community.

3.6 Results

Figure 3.2 displays the geographic distribution of counties with the most retweets. This graphic highlights a distinct geography of retweet activity over the 18-month study period. The counties that contain cities such as San Francisco, California; New York City, New York; Boston, Massachusetts; and the Northern Virginia portion of the greater Washington, D.C. metropolitan area are key hubs of retweet of activity. Interestingly, these locales represent important hubs on the Internet backbone [172] as well as well-known hubs of

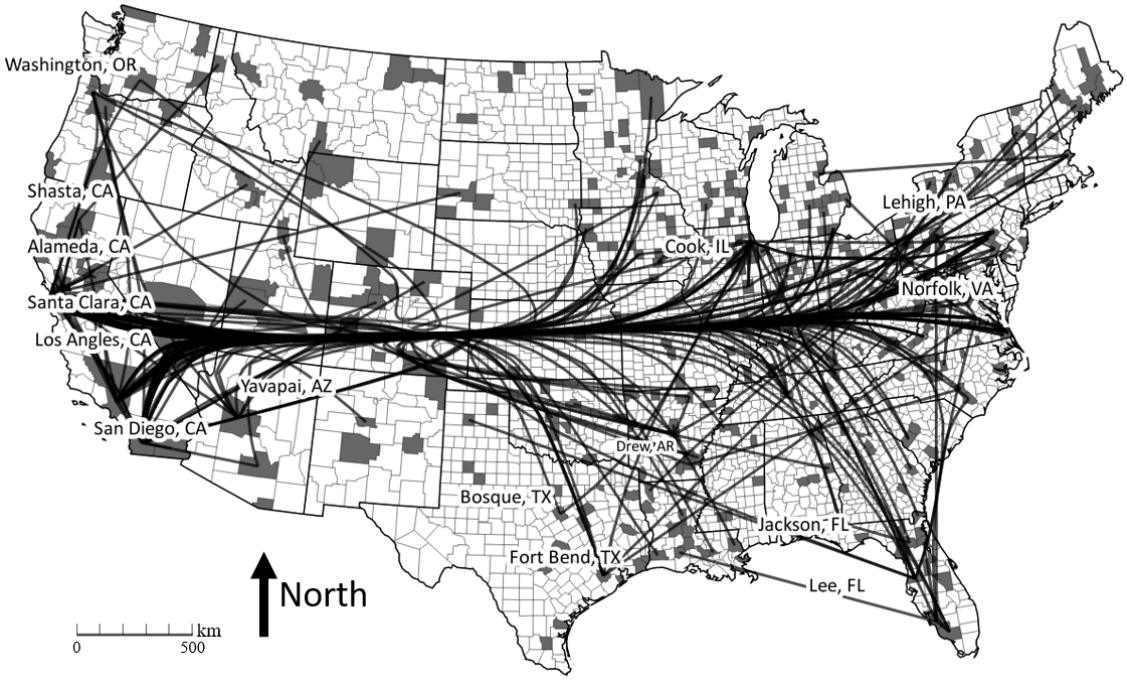


Figure 3.6: Community 3 Includes Counties such as San Francisco, CA, Henrico, VA and William, TN.

entrepreneurial activity [41, 145]. Other notably active counties, although somewhat less so than the aforementioned counties, include those pertaining to the Raleigh-Durham, North Carolina; Denver, Colorado; and Seattle, Washington metropolitan areas. The counties containing the Rustbelt cities of Detroit, Michigan and Cleveland, Ohio also appear to be somewhat active on Twitter. Several of the Great Plains states including North Dakota and South Dakota, as well as Western states such as Idaho and Wyoming are noticeably less active. This is not necessarily surprising given that Twitter data is more prevalent in urban and suburban communities [159], as well as the comparatively lower levels of broadband availability, in this portion of the country [16]. Aside from information about who is most and least active in producing Tweets pertaining to entrepreneurial activity, another notable aspect is the relatively static nature of retweet activity for the hashtags

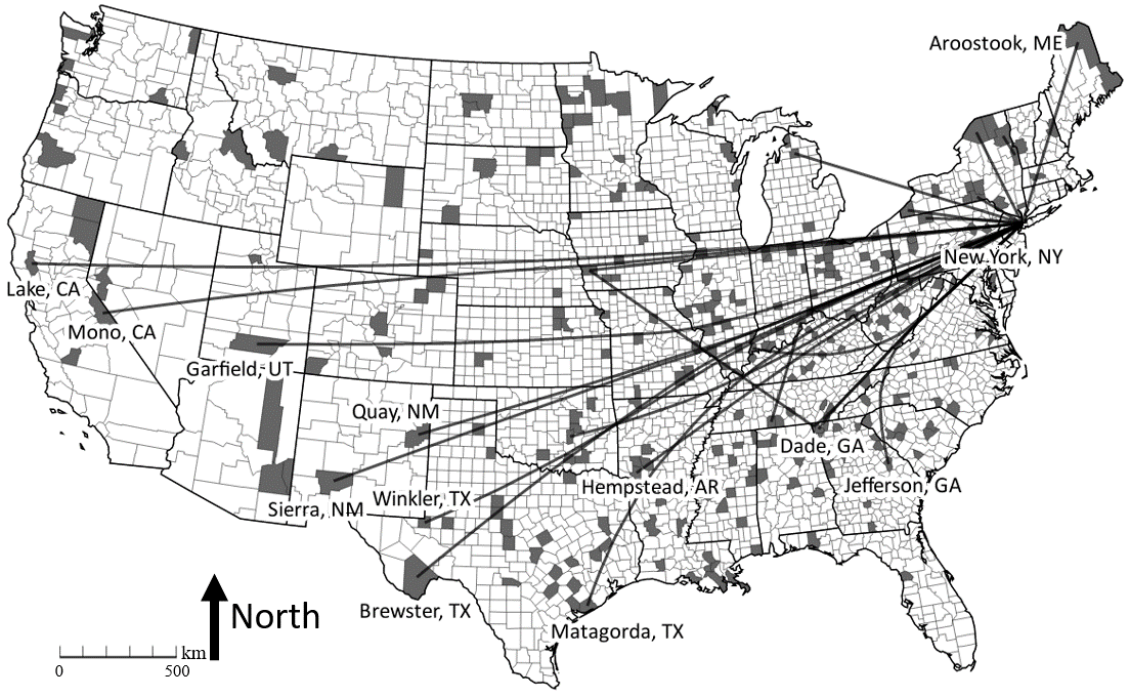


Figure 3.7: Community 4 Shows a Community Centered Around New York.

analyzed in this study. Prior to conducting the community analysis, a temporal inspection of retweets was done. The data was broken into six separate three month time intervals and inspected for consistency. While some locales, such as Fulton County, Georgia (Atlanta), have minor variations in terms of the magnitude of their Twitter activity, the maps of the three six-month sub-periods did not present dramatically different pictures of tweet activity. Figure 3.3 shows the normalized number of retweets by the population in the corresponding county. Many large metropolitan areas are also include. This pattern is also observed in the research of Mislove et al. [142]. Besides this, lots of sparsely populated areas are also highlighted. Many counties with research universities or natural amenities are also identified with lots of interactions.

Given the relatively static nature of the hierarchy of cities in terms of activity, this analysis will focus on networked interactions between counties over the whole 18-month time

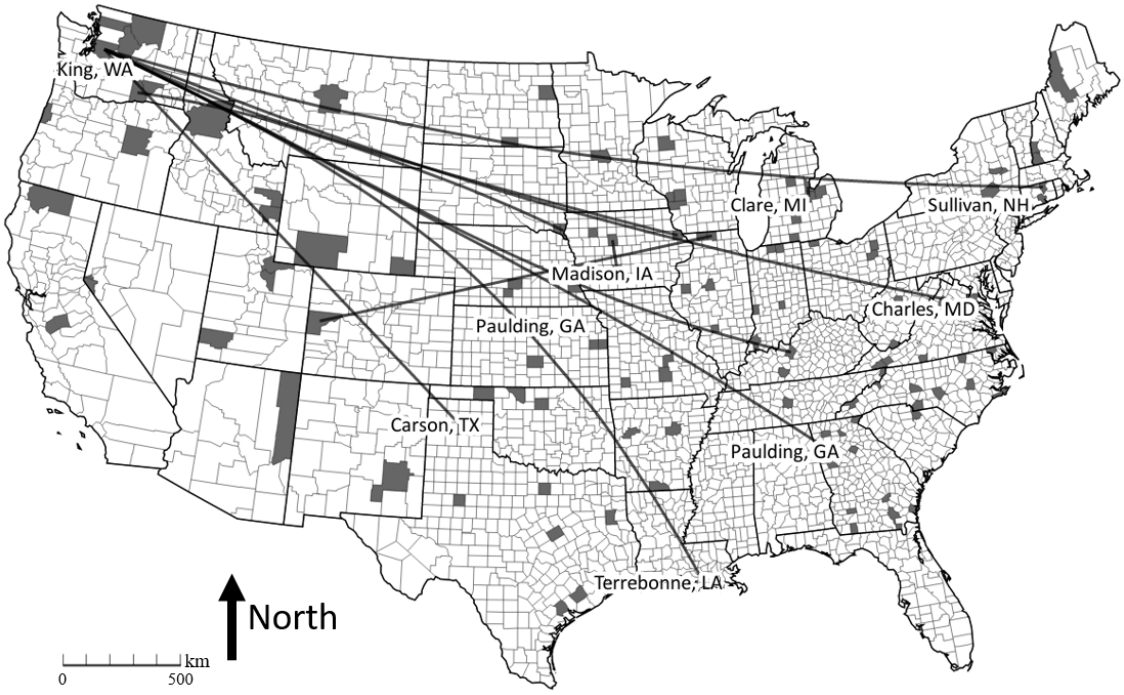


Figure 3.8: Community 5 Shows a Community Centered Around Seattle, WA.

period rather than analyzing each of the three 6-month sub-periods individually. Before moving on to a discussion of these geographies, one interesting aspect of Twitter activity pertaining to entrepreneurial activity (as defined in this study) are counties that represent key producers and consumers of information. Table 4 lists the top twenty cities in terms of both in-degree (consumers) and out-degree (producers), as indicated by network statistics computed for this analysis, and includes many of the counties containing the largest U.S cities in the country in terms of population, e.g., Atlanta (Cobb County, Georgia), Chicago (Cook County, Illinois), and Phoenix (Maricopa County, Arizona). One interesting county on this list is Mecklenburg County, North Carolina which is home to the city of Charlotte. The influential position of Charlotte in this type of Twitter network likely reflects its commitment to fostering entrepreneurship. In late 2012, Charlotte initiated an economic develop strategy centered on fostering high-growth entrepreneurship [173]. Given

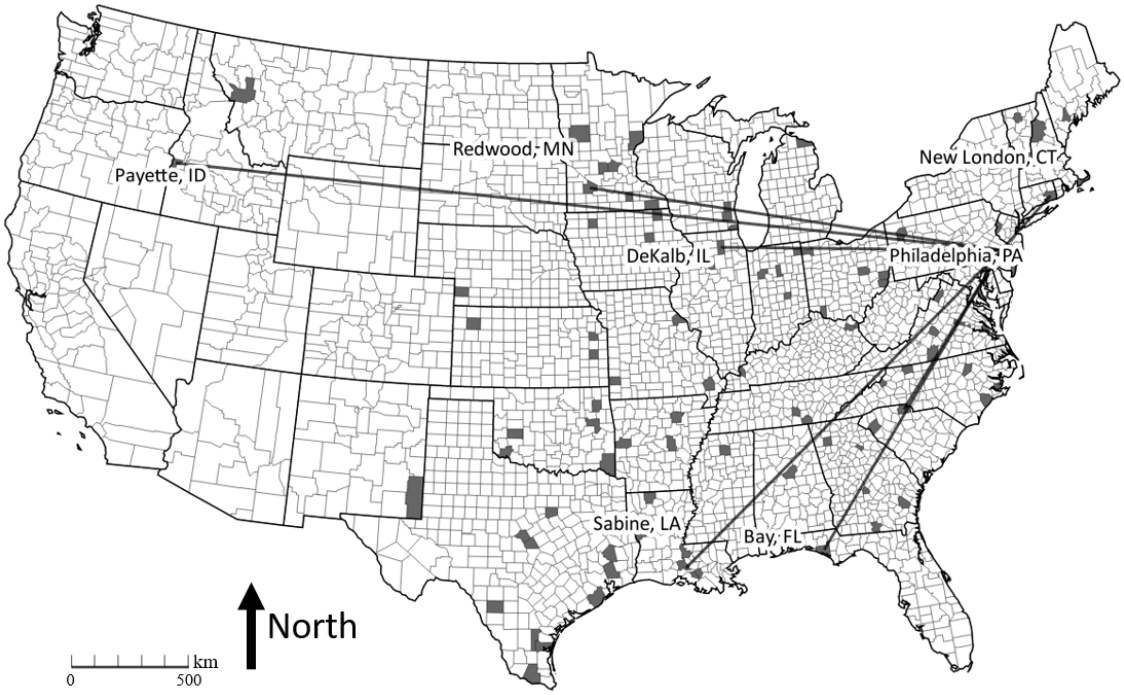


Figure 3.9: Community 6 Shows a Community Centered Around Philadelphia, PA.

this commitment to entrepreneurship there are several entrepreneurship networking organizations in the region, such as the Charlotte Chapter of the global Entrepreneurs Organization (EO) [174], and support organizations such as the Charlotte Regional Fund for Entrepreneurship (CRFE) [175].

3.6.1 Relational Geographies

The more interesting results of this analysis stem from the connectivity information provided by Twitter data, which can highlight the places that interact with one another most and least frequently. In an effort to provide more information about the locales that interact via Twitter most often using the hashtags indicating an interest in or participation in entrepreneurial activity, Table 3.4 lists the top centers of activity for the entire eighteen-month time interval. The top row contains the counties that are sources of the most retweet

activity. The column under each county in the top row lists a subset of the counties that retweeted information from these hub counties as a ranked list (in descending order). In the event of a tie, more than one county is included for a particular rank as necessary. Counties in this table are also categorized into one of four groups by geographic area: West, East, Midwest, and South. Based on these categories, it is somewhat easier to understand the physical distances over which these counties are interacting on Twitter. Some counties appear to interact primarily with counties that are geographically proximal. For example, New York, San Francisco, and Los Angeles all interact more frequently with counties that are located in their region. Several counties however do not follow this pattern and interact more frequently with counties outside of their region. Tweets coming from D.C., for example, are retweeted most often by counties on both coasts. Suffolk, Massachusetts tends to interact somewhat more often with counties on the west coast than on the east coast. San Diego also interacts more with counties on the opposite coast than counties on the west coast.

In an effort to better understand the interaction between counties, the Clauset-Newman-Moore algorithm was used to group counties into communities based on the intensity of interactions. As discussed in Section 3.1, communities are defined as those counties which consistently retweet each other, but have sparse interactions with counties outside of their community. Thus, the intensity of the interaction within the community is higher than its interaction with locations outside of the community. Figures 3.4-3.9 display the six major communities identified by this algorithm. By applying flow maps and filtering the network edges to show only flows between counties that have great than or equal to 5 retweets (for the ease of visualizing key nodes within these communities), some of the more interesting geographical properties of these communities can be illustrated. In this respect, communities 1-3 highlight places that interact in a transcontinental manner while communities 4-6

represent communities with one major node that appears to drive interactions with smaller locales.

Figure 3.4 (Community 1) has a distinct southern and eastern orientation to it. Key southern and eastern nodes on this network include Suffolk County, Massachusetts; Fairfax County, Virginia; and Cobb County, Georgia. Key nodes in Figure 3.5 (Community 2) include Marin County, California; Snohomish County, Washington; Norfolk County Massachusetts; and Mecklenburg County in North Carolina. This community is clearly transcontinental and includes a variety of locales on the east coast, on the west coast, in the south, and in the Midwest. Figure 3.6 (Community 3) includes counties such as San Francisco in California, Henrico in Virginia, and Williamson County in Tennessee. Finally, Figure 3.7 (Community 4) shows communities with distinct hub geographies consisting of tweets originating in New York, while Figures 3.8 and 3.9 highlight Seattle, Washington and Philadelphia, Pennsylvania, respectively, as key nodes in each of these communities. Although these communities are geographically nuanced, interestingly, they are rather similar in terms of their socio-economic and demographic characteristics. Table 3.5 displays the median values of several county characteristics including per capita income, population size, levels of broadband provision, and educational attainment. The definitions of these variables may be found in Table 3. Communities 1, 2, 3, and 5 are most similar in terms of their population size, number of broadband providers, and level of proprietors employment. Communities 4 and 6 appear to be more similar to one another, particularly with regard to their somewhat smaller populations, lower per capita income, and education attainment.

The goal of this study was to conduct an analysis of digitally mediated interactions using Twitter data about entrepreneurial activity via a methodology composed of existing research techniques including web-scraping, network analytics, community projection and flow maps. The analysis of these relational geographies revealed that although Twitter enables interactions across geographically distant locations, the highest intensity interactions

are regional in nature (Table 3.4). The community detection algorithm highlighted key nodes in transcontinental communities, as well as nodes (Philadelphia, Seattle, and New York) that exert their influence across large geographical distances. Despite these interactions across large geographical distances, interacting counties have similar socio-economic and demographic profiles suggesting social similarity is important to these interactions.

The results of the present study must be interpreted with some caution as algorithmic uncertainties inherent to data analysis may cause bias [17]. While precautions were taken to mitigate issues associated with geocoding Twitter information and user bias, and the results provide support for what is known about key nodes of entrepreneurship in a U.S. context, it is important to note that these results may be specific to the dataset at hand. It is also important to note that the lack of participation of users in more sparsely populated areas of the country limits the analysis of entrepreneurial networks via Twitter data to patterns and actors in urban locales. Research topics related to networks in peripheral areas are not appropriate given a lack of participation of geographically peripheral counties in this analysis. The age profile of Twitter users also suggests these data are not likely a good means of analyzing networks among older entrepreneurs. However, there are several reasons to believe that Twitter does hold promise for future analyses of entrepreneurial networks of younger entrepreneurs in urban locales.

	Count	**Explanations**
Number of unique users in the dataset collected globally	297,639	
Number of unique users located in the U.S.	89,914	30% of all global users provided usable information for the U.S.
Number of users with profile data	**89,411**	Usable profile information that can be geocoded to U.S. counties
Number of users with Global Positioning Data (GPS)	1,744	
Number of users with both profile data and GPS data located in the US	1,241	This is a subset of the 89,411 users with profile data.
Number of users with only GPS Data	**503**	Users with no profile information but GPS information.
Number of Users where GPS Data does not match profile data	217	1,241 users had profile and GPS data; of these users, 217 had GPS data and profile data geocode to different counties.
Number of U.S. based users with multiple profile entries (e.g., Madison/Milwaukee/Chicago)	314	User profiles with multiple entries. First location used to geocode user to county.
Total Number of Unique Users Geolocated in U.S.	**89,914**	

Table 3.2: User Summary

Rank	Out-Degree (Producers)	In-Degree (Consumers)
1	New York, NY	San Francisco, CA
2	Washington D.C.	New York, NY
3	San Francisco, CA	Travis, TX
4	King, WA	Washington D.C.
5	Cobb, GA	Los Angeles, CA
6	Los Angeles, CA	Cook, IL
7	Philadelphia, PA	Cobb, GA
8	San Diego, CA	Norfolk, MA
9	Santa Clara, CA	Suffolk, MA
10	Suffolk, MA	Santa Clara, CA
11	Cook, IL	King, WA
12	Maricopa, AZ	Mecklenburg, NC
13	Orange, FL	San Diego, CA
14	Travis, TX	Yolo, CA
15	Henrico, VA	Cumberland, PA
16	Dallas, TX	Philadelphia, PA
17	Hudson, NJ	Dallas, TX
18	Davidson, TN	Harris, TX
19	Kings, NY	Adams, CO
20	Mecklenburg, NC	Miami-Dade, FL

Table 3.3: Key Producers (Out-Degree) and Key Consumers (In-Degree) of Tweets

Top Centers of Twitter Activity	The counties with the most interactions with the center (Top Row)										
Rank	**New York, NY**	**San Francisco, CA**	**Los Angeles, CA**	**Cobb, GA**	**Washington D.C.**	**Cook, IL**	**Travis, TX**	**San Diego, CA**	**Suffolk, MA**	**Mecklenburg, NC**	
1	San Francisco, CA	New York, NY	San Francisco, CA	San Francisco, CA	New York, NY	New York, NY	San Francisco, CA	San Francisco, CA	San Francisco, CA	New York, NY	
2	Washington D.C.	Los Angeles, CA	New York, NY	New York, NY	San Francisco, CA	San Francisco, CA	Houston, TX	New York, NY	New York, NY	San Francisco, CA	
3	Los Angeles, CA	San Diego, CA	Santa Clara, CA	Fulton, GA	Cook, IL	Washington D.C.	New York, NY	Albany, NY	Mecklenburg, NC	Ashtabula, OH	
4	Cook, IL	Santa Clara, CA	Washington D.C.	Los Angeles, CA	Los Angeles, CA	Denton, TX	Los Angeles, CA	Maricopa, AZ	Santa Clara, CA	Suffolk, MA	
5	Mecklenburg, NC	Travis, TX	San Diego, CA	Washington D.C.	Fairfax, Fairfax City + Falls Church, VA	Santa Clara, CA	Mecklenburg, NC	Los Angeles, CA	Middlesex, MA	Dallas, TX	
6	San Diego, CA	Mecklenburg, NC	Cook, IL	Cook, IL	Cobb, GA	Los Angeles, CA	King, WA	Norfolk, MA	Washington D.C.	Denton, TX	
7	Suffolk, MA	King, WA	Travis, TX	Philadelphia, PA	Suffolk, MA	Cobb, GA	Washington D.C.	Washington D.C.	Cook, IL	Weber, UT	
8	Cobb, GA	Suffolk, MA	Cobb, GA	Miami-Dade, FL	Philadelphia, PA	Suffolk, MA	Cook, IL	Miami-Dade, FL	Los Angeles, CA	Travis, TX	
9	Kings, NY	Cobb, GA	King, WA	Suffolk, MA	King, WA	Philadelphia, PA	Suffolk, MA	Cook, IL	Cobb, GA	Cumberland, PA	
10	Philadelphia, PA	Cook, IL	Orange, CA	Norfolk, MA	Sullivan, NH	Travis, TX	Cobb, GA	King, WA	King, WA	Cobb, GA	
11					San Diego, CA	King, WA					
	West		**Midwest**		**South**			**East**			

Table 3.4: Key Hubs of Twitter Activity

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Population	51,581	75346	59981	36729	50755	43363
Per Capita Income	35,776	37,635	37,107	35,137	37,504	36,358
Median Age	39.8	39.3	39.7	40.2	39.65	40.1
Percent Bachelor's or Higher	20%	22%	20%	17%	21%	18%
Percent White	89%	86%	89%	91%	90%	91%
Percent Black	3.7%	4.3%	2.6%	2.5%	2.2%	2.7%
Percent American Indian	0.31%	0.31%	0.35%	0.30%	0.37%	0.35%
Percent Asian	0.76%	1.0%	0.75%	0.52%	0.84%	0.67%
Proprietor's Employment	6127.5	8895	7085	4526	6597.5	5751
Number of Broadband Providers	78.5	108	87	57	85	70
Natural Amenity Index	0.020	0.09	-0.12	-0.10	-0.28	0.055
Number of Counties	446	445	391	285	126	98
Number of Users	15,565	37,527	21,712	9,777	2,627	1,848
Number of Retweets	7,957	22,556	14,845	16,209	947	651

Table 3.5: Cluster Profiles

Chapter 4

APPLICATIONS OF COMMUNITY DETECTION IN VISUAL ANALYTICS

In the last chapter, community detection is applied in social network analysis. Many networks present community structures. A closely connected community might imply faster and more frequent information exchange among the members. Thus, the communities are defined as a group of nodes that are densely connected with other nodes within the same communities while they are sparsely connected with other nodes in the network. In this chapter, different network community detection algorithms and their applications in visual analytics are compared. Many measures can be used to identify the communities, such as node degrees, clustering coefficient, betweenness, and centrality. The focus is the algorithms based on modularity optimization. Besides the network structure, geographical information can also be integrated into modularity optimization process.

4.1 Definitions

In a network, modularity (Q) quantifies the community strength by subtracting the expected fraction if edges were distributed at random from the fraction of edges within communities. The value of modularity is between $[-\frac{1}{2}, 1]$. The positive modularity implies that the nodes within the assigned groups are more likely to be connected than random. The lower bound is achieved from any bipartite graph with a canonic clustering (C_0, C_1). As there are no edges within the communities and all edges are between the nodes with different clusters, the modularity is $-\frac{1}{2}$. The upper bound is constructed from a graph with no edges and each cluster only contains one corresponding node. In this case, each cluster is inseparable and cannot be merged through any edge connection. Formally, modularity can

be defined as:

$$Q = \frac{1}{2m} \sum_{1 \leq i,j \leq n} [A_{ij} - \frac{k_i k_j}{2m}] \delta_{C_i, C_j} \tag{4.1}$$

$$\delta_{a,b} = \begin{cases} 0 & a \neq b \\ 1 & a = b \end{cases} \tag{4.2}$$

where m is the number of edges in the network, n is the number of nodes in the network, A is the adjacency matrix, k_i is the degree of node i, δ is Kronecker delta, and C_i is the community where node i is assigned.

From the above definitions, maximizing modularity can be used to get the community structure of networks. However, optimization of modularity is an NP-complete problem [176]. Greedy algorithms can approximate the optimization process (such as the Girvan-Newman algorithm [97], the CNM algorithm [171], the Wakita-Tsurumi algorithm [177], and the Louvain method [178]). In the next sections, several greedy algorithms and their variants are used. As they would use dendrogram-like procedures, the hierarchical community structure is extracted.

4.2 Hierarchical Structure in Community Detection and Visual Analytics

Most greedy algorithms start from n clusters and merge the clusters in a tree structure. Besides the final clustering results, the intermediate results can be also used to show the inner structure within a cluster. Here the CNM algorithm is used as an example. In each iteration, the two communities that can contribute maximum increase of modularity are merged. From the definition of modularity, the "closest" two clusters are merged in each iteration. In Chapter 2, the visualizations of geographical networks have high visual complexities because of the size of the network. The sub-dendrogram trees generated during the modularity optimization procedure can be used to perform filtering to reveal the inner structures of the clusters.

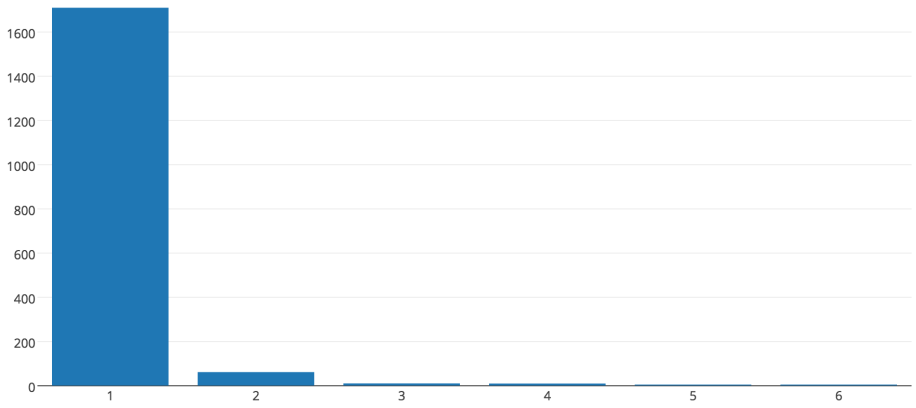


Figure 4.1: Size Distribution of Smaller Clusters Merged in CNM Algorithm.

One of the most straightforward methods is to visualize the dendrogram and the links between nodes. One example of tree-based network visualization is Latour [179]. However, this method suffers when the network is large. First, all observations are leaf nodes in dendrogram trees. That could lead to very wide plots when the number of n is large [180]. Second, tree-shaped network visualizations suffer from visual clutter and insufficient use of space. Especially for the CNM algorithm, it is found that small clusters in the size of 1 or 2 nodes that were merged into large clusters quite frequently, and the largest height for the community is 517. Figure 4.1 shows the distribution of the size of smaller communities were merged. To visualize a network with a tree-shaped layout, two layout structures can be used: the planar layout and the circular layout. An interactive prototype is proposed to explore the community detection structure based on the modularity optimization procedure. When the user selects a community, the corresponding range of modularity of the community is highlighted. Then the user can select a modularity value and the sub-community generated at that modularity is shown on the corresponding map. The user can explore the community construction process by selecting a node. To find when a single county got merged into its corresponding cluster, the user can click on the map to specify a county. The nodes under that corresponding modularity are highlighted. Figure 4.2 shows the pro-

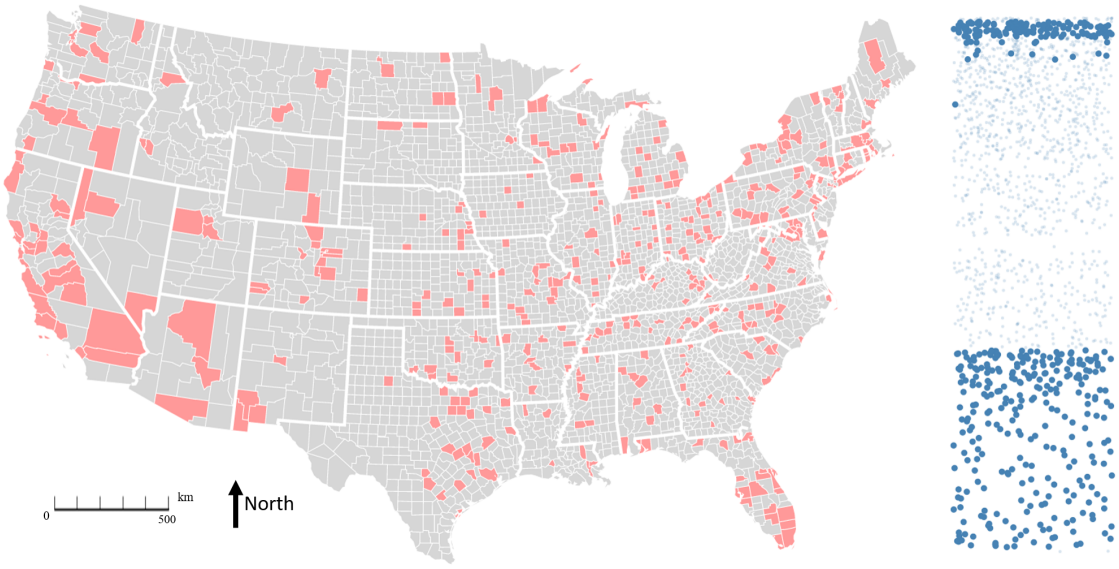


Figure 4.2: The Visual Analytics Framework for Community Structure Detected by CNM Algorithm. Community 2 is Used as an Example Here.

totype interface. From the density of the nodes on the right side, it can be observed that the modularity increases linearly during the optimization, which also matches the observations by previous experiments [171, 181].

The construction process can be analyzed through exploring the modularity values and the corresponding nodes in the community. Community 2 is used as an example to illustrate the analytics procedure in Figure 4.2. First, the modularity distribution is shown on the right. It can be observed that there are two major clusters during the modularity procedure. Next, the user can select counties on the map or nodes on the right to show the nodes to the corresponding modularities. The visualization shows that there are some neighboring coastal California and Florida counties within this community. The first interesting pattern is that in the first stage of the community construction, these neighboring counties have been merged into the community as shown in Figure 4.4. Similar patterns in community 3 are found. Figure 4.3 shows the nodes distribution of community 3. Three major stages

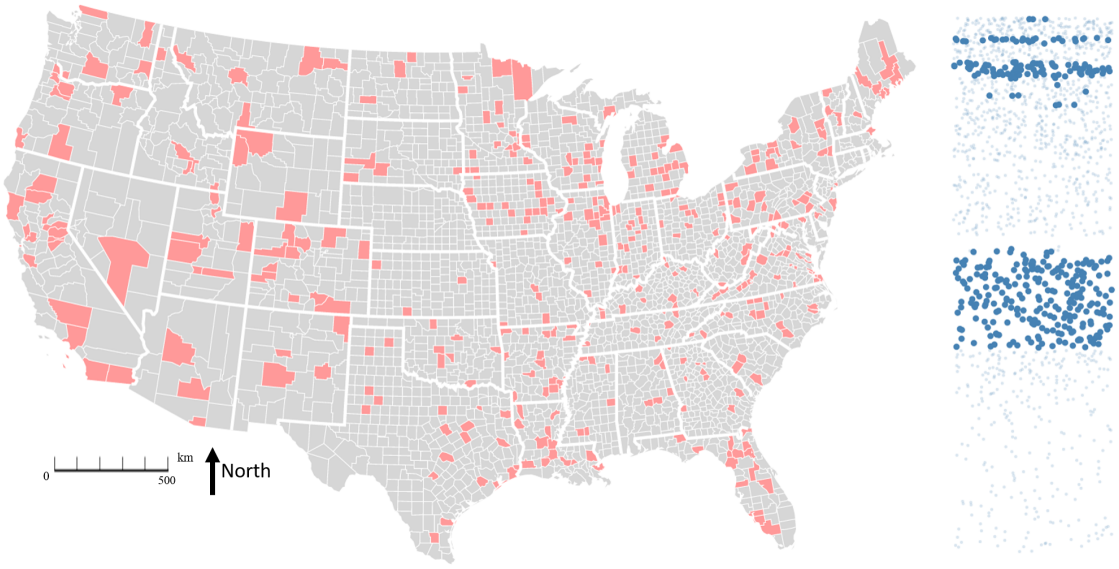


Figure 4.3: Nodes and the Modularities During the Construction of Community 3. Three Major Stages can be Identified.

can be identified during the modularity increase process. Figure 4.5 show these three major stages during the optimization. It shows that the neighboring counties are merged into the community at about the same time. In the CNM algorithm, the two communities which contribute maximum increase of modularity are merged. This heuristic strategy implies that the neighboring counties contribute more to the increase of modularity, which means that there are more interactions between neighboring counties.

4.3 Integrating Geographical Information with Community Detection

Louvain method is another greedy modularity optimization algorithm for community detection. Besides the time and space complexity, the Louvain method also generates higher modularities according to previous experiments [178]. In this section, the hierarchical structure of the communities with linked visualizations is analyzed. Then three variations of Louvain modularity are compared in the following section.

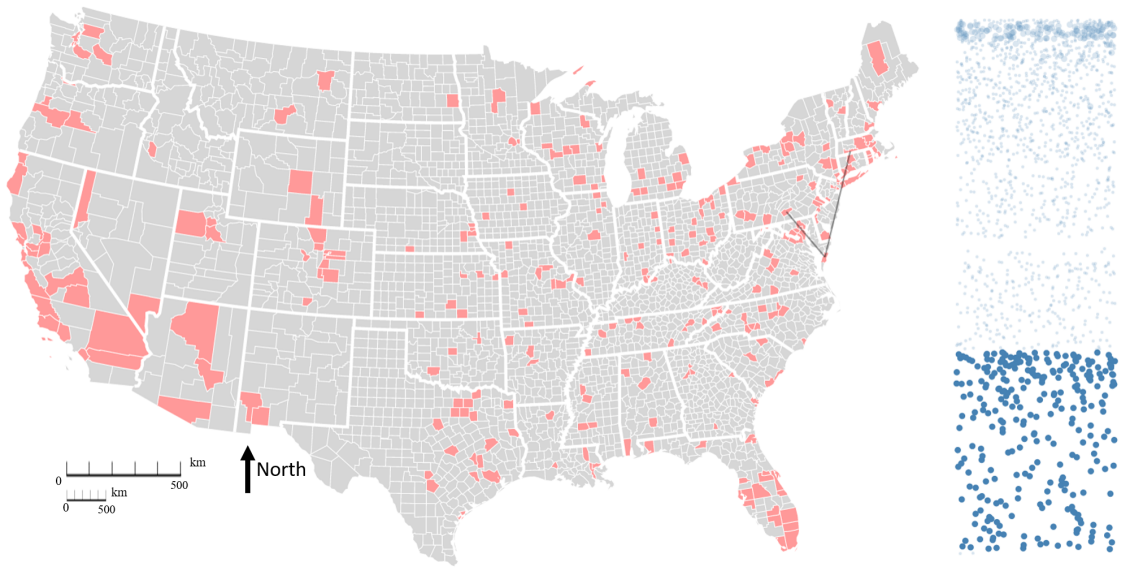


Figure 4.4: Most Neighboring Counties in California and Florida have been Merged into the Community 2.

Louvain modularity optimizes the time complexity with a hierarchical local optimization procedure. At first, all nodes are assigned into their own communities. Each iteration includes two phrases. The first phrase checks each node c_i for its neighbors (for example, c_j), and tests the increase of modularity defined as:

$$\Delta Q = \left[\frac{\sum_j + 2k_{i,j}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)\right] - \left[\frac{\sum_j}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{4.3}$$

where $\sum_j$ is the sum of all weights of edges inside of community c_j which is being tested, and $\sum_{tot}$ is the sum of all weights of the edges from c_i to c_j.

If there is at least one neighbor c_j which can provide a positive increase of modularity, then c_i will be merged into the neighbor which provides the largest increase. In the second phase, the graph is aggregated by grouping all nodes in the community into a single node. Then this induced graph is used in the next iteration until the modularity cannot be increased anymore.

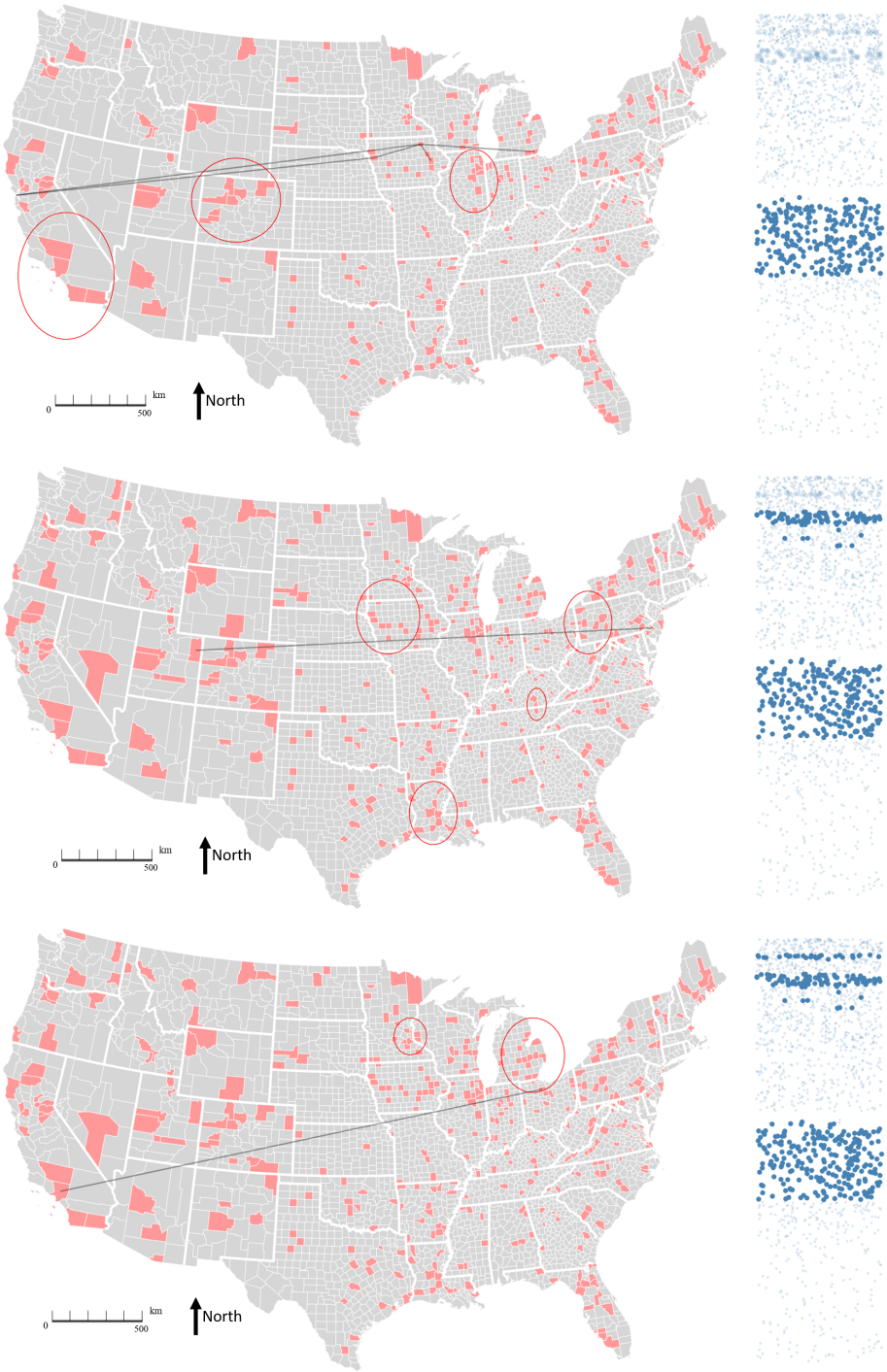


Figure 4.5: Three Stages of Merging Nodes into Community 3. The Neighboring Counties are Highlighted in the Images.

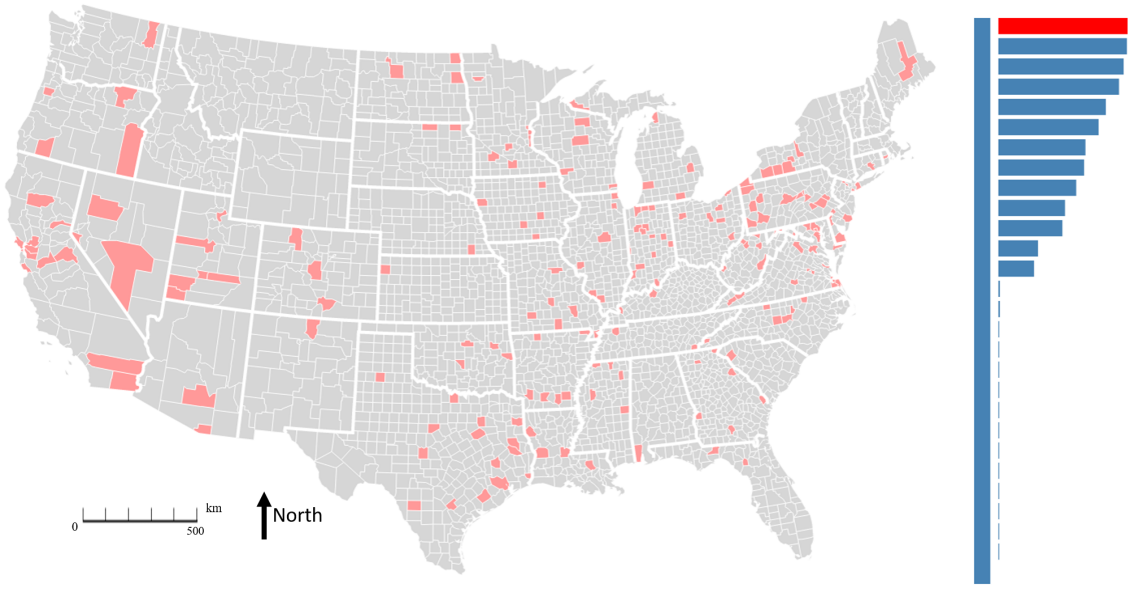


Figure 4.6: Visual Analytics for Hierarchical Clustering.

To analyze this aggregation procedure, a visual analytics method similar to the previous analysis of CNM algorithm can be used. The communities are sorted in the order of their sizes. Upon the selection of one community, the sub-communities from the previous iteration are shown. Figure 4.6 shows the system interface for analysis of modularity algorithms. The counties with more interactions can be identified by exploring the sub-communities.

Besides the virtual communities extracted from the network, the geographical background information is also integrated to examine the importance of spatial distance in the social network interactions. Based on the highly efficient optimization procedure of Louvain method, there are several variations to take geographical distance into consideration [182–184]. The results between CNM [171], Louvain[178], Louvain-SN, SNIC Heuristic [183], and Louvain-D [184] are compared.

Louvain-SN (Spatial Near) algorithm scales modularity with the aggregated distance between the node locations and the centroid of the corresponding community [183]:

$$M_{SN}(C,\delta) = \frac{1}{2m}\left(\frac{\sum_{c\in C}\sum_{i,j\in c} w_{ij} - \frac{k_i k_j}{2m}}{1 = \sum_{c\in C}\underset{i\in c}{\text{agg}}\,(\frac{d(i,x_c)}{\sigma})^2}\right) \tag{4.4}$$

where x_c is the centroid of the nodes in the community c, σ is the scaling parameter and $d(i,j)$ is the geographical distance between node i and j. The importance of distance in the model is scaled down as σ increases. Specifically, Equation 4.4 is equivalent to Equation 4.2 when $\sigma \to \infty$. In this experiment, the *max* function is used for the distance aggregation.

The SNIC (Spatial Near, Iterative Constraining) Heuristic algorithm extends the Louvain-SN method by using a variable scaling parameter and more strict spatial constrictions [183]. In the first phrase of each iteration, a node can be merged only if its distances to all nodes in the target community are within the constraint. σ is changed into the maximum distance within communities. The initial σ is reached with the initial partition generated from the first iteration of Louvain-SN.

Louvain-D extends the modularity by scaling distance with a kernel function [184]:

$$M_{dist}(C,\sigma) = \frac{1}{2m}\sum_{c\in C}\sum_{i,j\in c} w_{ij} - P_{ij} \tag{4.5}$$

$$P_{ij} = \frac{\hat{P}_{ij} + \hat{P}_{ji}}{2} \tag{4.6}$$

$$\hat{P}_{ij} = \frac{k_i k_j f(d(i,j)/\sigma)}{\sum_{q\in V} k_q f(d(q,i)/\sigma)} \tag{4.7}$$

where f is the kernel function. The Gaussian function is used in this experiment:

$$f(u) = e^{-u^2} \tag{4.8}$$

In these experiments, the scaling parameter δ is set as 3000 kilometers.

It is found that that the results from methods based on Louvain modularity disagree with the results from CNM algorithm. First, the results are compared visually with the visual analytics interfaces. It is found that some clusters partially match the sub-modularity

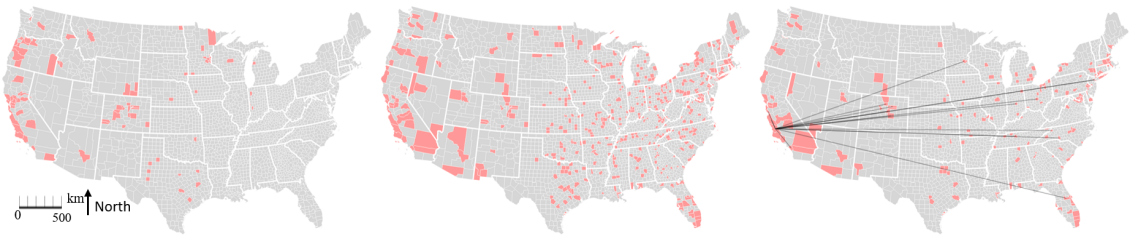


Figure 4.7: Comparison Between Communities Discovered by the Louvain-SN and CNM algorithms.

discovered in the CNM algorithms. Figure 4.7 shows an example of the comparison between clusters from different algorithms. The left image is a community discovered by the Louvain-SN algorithm. The middle image is community discovered by the CNM Algorithm. The right image shows a sub-community found using the CNM algorithm during runtime. This figure shows that although the counties in southern coastal California match with the early stage results from CNM, there are still many differences between these clusters. Although algorithms like Louvain-SN integrate geographical distance in the optimization procedure, is still shows that the community of Louvain-SN misses lots of counties that are spatially near to the community. However, it is still very complicated to analyze such difference using a side-by-side visual comparison. In the next section, the communities are compared quantitatively.

4.4 Comparisons Between Different Community Detection Algorithms

These modularity variations can be used to evaluate the detected communities in the geographical context [183, 184]. Figure 4.8 compares modularities. A surprising result is that Louvain-D suffers in all of the three modularities. Louvain-SN and Louvain-SNIC outperform other methods in the respect of their target function. From the previous visualizations, it can be observed that neighboring counties can be often grouped into the same community in the early stage, which implies that they have more interactions than other

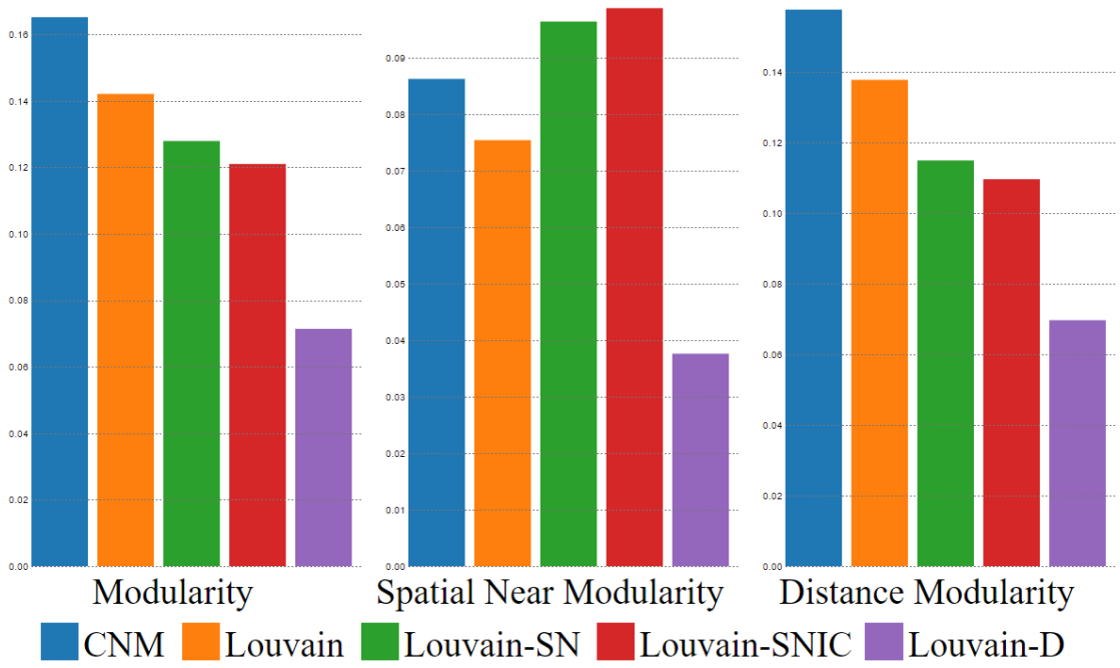


Figure 4.8: Community Detection Comparison With Modularity Variants.

counties in the same community. Although adding geographical coordinates does not increase the resulting modularity, it helps to detect the regional communities such as the one shown in Figure 4.7 (left).

Besides the modularities, it is neccesary to investigate the difference between these results for details. To compare the results from different algorithms, metrics based on counting pairs [96], set matching [185] and variation of information [186] can be used. In the following discussions, two clustering results are defined as C and C'.

The first metric is based on counting pairs. This method counts the number of node pairs (p,q) in the following scenarios:

- N_{11} counts the pairs that are under the same cluster both in C and C',
- N_{00} counts the pairs that are under different clusters both in C and C',
- N_{10} counts the pairs that are under the same cluster in C but not in C',

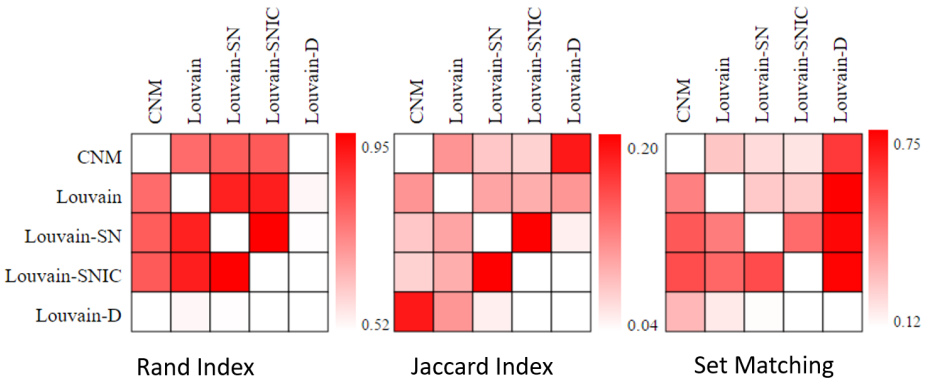


Figure 4.9: Clustering comparison with three metrics.

- N_{01} counts the pairs that are under different clusters in C and not in C'.

It is simple to find that $N_{11}+N_{00}+N_{10}+N_{01} = n(n-1)/2$. Based on these counts, Wallace proposes two asymmetric metrics[96]:

$$W_I(C,C') = \frac{N_{11}}{\sum_{k=1} n_k(n_k-1)/2} \tag{4.9}$$

$$W_{II}(C,C') = \frac{N_{11}}{\sum_{k=1} n'_k(n'_k-1)/2} \tag{4.10}$$

where n_k is the size of community k in C and n'_k is the size of community k in C'.

Fowlkes and Mallows combines these two metrics with geometric mean [187]:

$$\mathscr{F}(C,C') = \sqrt{W_I(C,C')W_{II}(C,C')} \tag{4.11}$$

This metric is a useful baseline to compare communities if there are many clustering results. However, this metric has no fixed lower and upper bounds. To get a uniform metric, adjusted Rand index can be applied [188, 189]:

$$\mathscr{R}(C,C') = \frac{N_{11}+N_{00}}{n(n-1)/2} \tag{4.12}$$

Where $\mathscr{R} \in [0,1]$ and $\mathscr{R} = 1$ when $C = C'$. Figure 4.9 (left) shows the Rand index results between these clustering methods.

Other set overlapping metrics, such as the Jaccard index, can be extended into the community detection results [190]:

$$\mathscr{F}(C,C') = \frac{N_{11}}{N_{11}+N_{01}+N_{10}} \tag{4.13}$$

Figure 4.9 (middle) shows the Jaccard index between these algorithms. The major difference between the Jaccard index is from N_{00}. From the difference, it can be seen that Louvain-D is not very likely to merge counties in different clusters into the same community. It is very likely that Louvain-D just merges counties, which could result a large decrease of modularity. This is very common in the procedure of hierarchical clustering. So, one more metric is needed to analyze this scenario.

If there are split clusters $c_3 = c_1 \cup c_2$ where $c_1, c_2 \in C_1$ and $c_3 \in C_2$, large values of N_{10} and N_{01} can be found, even though these two partitions are actually quite similar. Another issue is that this method cannot be scaled to larger tuples, such as triads, easily. Thus it can be considered using the second metric which is based on set matching. The best match $c'_i \in C'$ with the maximum number of overlapped nodes can be found for each cluster $c_i \in C$. Then the ratio of the overlapped elements as the similarity metric can be used [185]:

$$\mathscr{H}(C,C') = \frac{1}{n} \sum_{c'_i = match(c_i)} n_{ii'} \tag{4.14}$$

where $n_{ii'}$ is the overlapped number of nodes between clusters c_i and c'_i. Figure 4.9 (right) shows the set matching difference. From the right images, it can be seen that the column for Louvain-D has very high values. That implies that the communities got merged into the communities in Louvain-D. Although modularity is an important metric to evaluate the detected communities, global modularity is still need to be carefully used. The local community structure can be quite different from the results yielded from algorithms based on global modularity optimization. In some cases, the local modular structures can be more important [191]. So another method can be analyzing subgraphs with filters such

as geolocation fileters. The visual analytics tools provide a data exploration procedure for model exploration so that the user can select the best algorithm and settings for the task.

4.5 Temporal Dynamics of Social Networks

Another task of social network analysis is to analyze the network dynamics. Most dynamic social network analysis methods divide the network into snapshots taken over time. The evolution behaviors include new node creation, node removal, and structural changes, such as new edges or removed connections. To analyze the properties of the graph, the methods in the previous sections need to compute over all paths in the graph, which is expensive. To solve this issue, the nodes in a graph can be projected in to an k-dimension space. The most intuitive projection is to use the adjacency matrix, which has the dimension of n. Then the closeness between the nodes can be measured with Euclidean distance. However, the added dimensions also makes the computation more expensive. It was shown that any finite graph can be projected to three-dimensional space through graph embedding [192]. To reduce the number of dimensions, graph embedding is a popular method for graph dimension reduction [193]. This method can also reveal the weak connections. Skillicorn et al. propose a spectral embedding method for dynamic social network analysis [194]. This method embeds the network snapshots into the same k-dimensional space so that the patterns movement patterns of the nodes can be analyzed. First, the aggregated adjacency matrix A_t in time frame t is composed with adjacency matrices W_t and the adjacency matrix W_{t-1} in the previous time frame:

$$A_t = \begin{cases} W_t, & t = 1 \\ (1-\alpha) * W_t + \alpha * A_{t-1}, & t > 1 \end{cases} \tag{4.15}$$

The random walk matrix R_t is generated from A_t by dividing each row by the corresponding row sum. To avoid the nodes getting trapped in the local circle, a small distur-

bance ε is introduced:

$$R_t^{new} = (1-\varepsilon) * R_t^{old} + \frac{\varepsilon}{n} * J \tag{4.16}$$

where n is the number of nodes and J is a $n \times n$ all-ones matrix.

Next, the c time stamps are stacked into a new random walk matrix M:

$$M = \begin{pmatrix} (1-\beta)R_1 & \cdots & \frac{\beta}{c-1} * I & \cdots & \frac{\beta}{c-1} * I \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\beta}{c-1} * I & \cdots & (1-\beta)R_2 & \cdots & \frac{\beta}{c-1} * I \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\beta}{c-1} * I & \cdots & \frac{\beta}{c-1} * I & \cdots & (1-\beta)R_c \end{pmatrix} \tag{4.17}$$

Three parameters are used in this model. α reflects the smoothing from the previous time frame. β defines how likely it is that the network can be aligned to its previous state. β should be larger than ε to ensure that local subgraph structure cannot dominate temporal changes. α, β and ε are set as 0.5, 0.3, and 0.2 in the following experiments.

Then the Laplacian matrix L can be constructed as:

$$L = I - \frac{\Pi^{1/2} M \Pi^{-1/2} + \Pi^{-1/2} M' \Pi^{1/2}}{2} \tag{4.18}$$

where I is an identity matrix. Π is a diagonal matrix containing stationary distributions of the nodes in M. Let λ_i and g_i be the ith eigenvalue and corresponding eigen vector. The embedding vector f_i is adjusted from g_i by the stationary distribution:

$$f_i = \Pi^{-1/2} g_i \tag{4.19}$$

Figure 4.10 shows the overall embedding in 3D space. Besides a few outliers, most nodes were clustered together. Then the focus is to examine the dynamics of nodes in the position of graph evolution. Figure 4.11, 4.12 and 4.13 shows the temporal dynamics of the leading nodes in their corresponding clusters. In Figure 4.11, Orange County, CA connects two sub-clusters. This implies that the impacts of this county shifts from one subcluster to

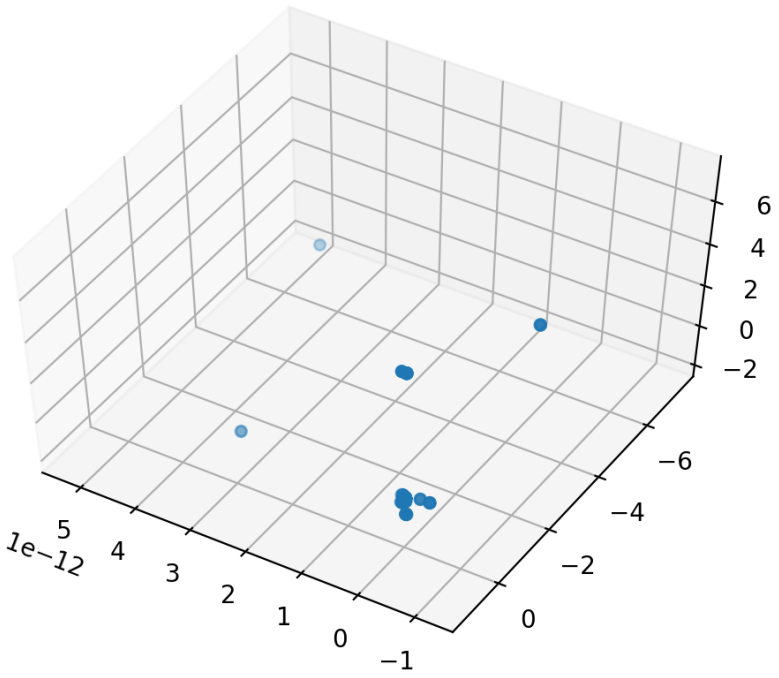


Figure 4.10: Laplacian Embedding of the 6 Phrase Dynamic Network With Edges Omitted.

another. In Figure 4.12, Los Angles County stays still. This implies that its impact is stable during the temporal evolvement. In Figure 4.13, Philadelphia County is "escaping" from its cluster and it is further that the other examples. This implies that this leading node has less impacts to the other nodes in its cluster and the impact is decreasing.

4.6 Resilience in Global Trade Network

In network analysis, another issue to consider is the robustness against attacks. In global trade networks, the attack can be sudden increase or decrease of supplies and change of polices. Matisziw et al. propose a analytics model for network robustness with edge removal testings [195]. Network resilience is defined as the ability to maintain an acceptable service

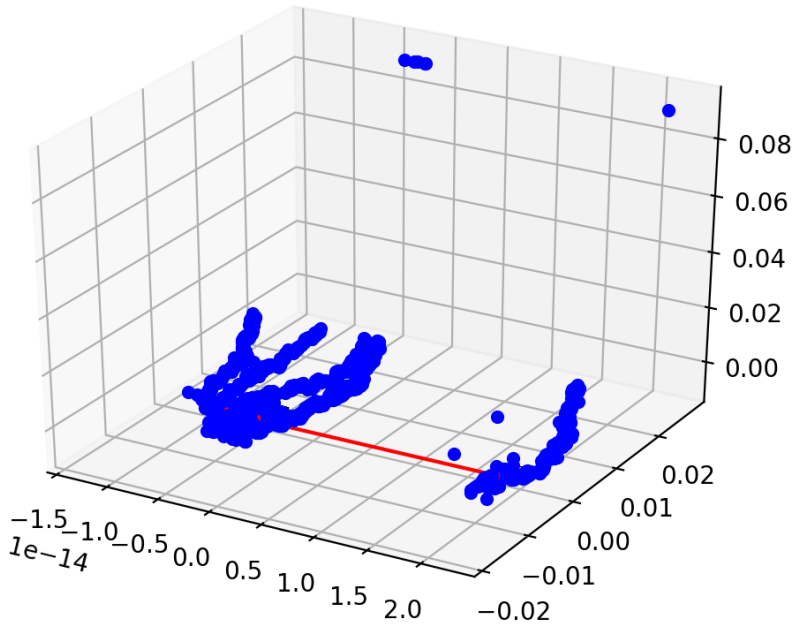


Figure 4.11: Embedding Results of Cluster 2. Orange County, CA is Highlighted in Red.

during failures. Redundancy is a polular strategy to increase the resilience of a network. To measure the redundancy, one of the most straightforward methods is to measure the dependency to the key nodes. Chen et al. propose a network optimization algorithm which uses first eigen value as vulnerability score [196]. Alenazi and Sterbenz propose flow robustness, which models the robustness with size of network components [197]. Malliaros et al. propose a generalized robustness index based on subgraph centrality [198]. Benzi and Klymko propose the resilience measure with total connectivity [199]. In the global trade network analysis system, these resilience metrics are used to measure the stability upon sudden changes in global trade.

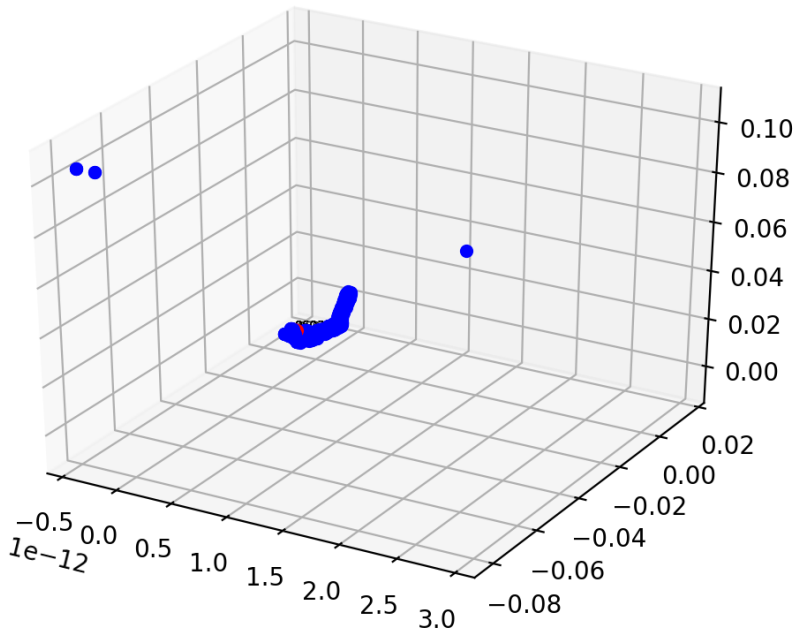


Figure 4.12: Embedding Results of Cluster 3. Los Angles County, CA is Highlighted in Red.

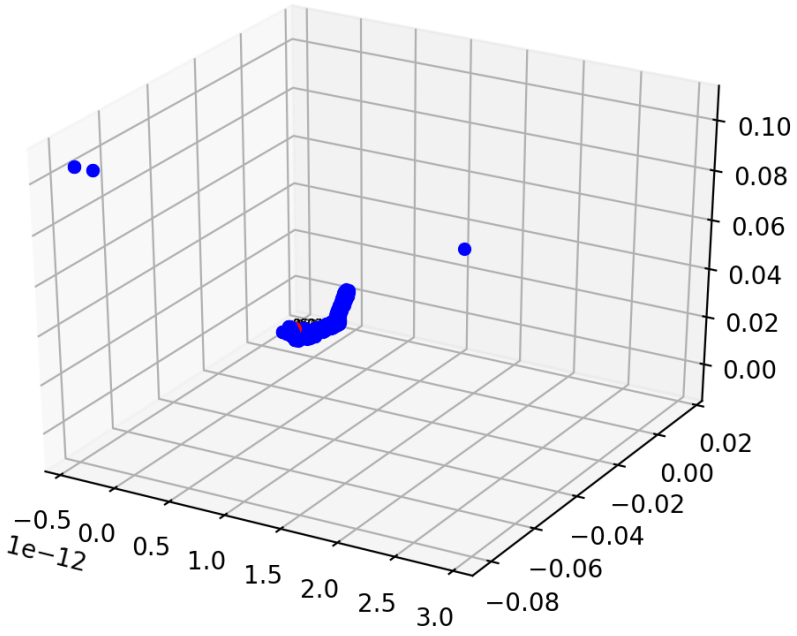


Figure 4.13: Embedding Results of Cluster 6. Philadelphia County, PA is Highlighted in Red.

Chapter 5

PHYSICAL NETWORK HOTSPOTS

The second integration between network and spatial data is to integrate physical networks and data points. This thesis presents techniques to abstract the topology of geospatially constrained nodes and edges, such as urban street networks, and use the networks as the space where the problems are embedded. In this thesis, urban street networks are used as the complimentary information in hot spot identification and territory estimation. This thesis argues that for many event based data types, the linking of network properties to the aggregation model can reveal hidden structure in the resultant analytic representation of the data and serve as a novel means for both visualization and analysis. This thesis proposes modifying the kernel density bandwidth, forgoing fixed spatial distances, and, instead, uses network properties (such as the speed limit of roads) as constraints on the bandwidth distance. Using the network based density estimation, this thesis devises a methodology for extracting hotspots based on their geographic network boundaries. These extracted territories are then analyzed for overlap to enable the identification of areas with compound risks (i.e., areas where several unique gangs are active or areas where multiple unique types of crime are likely to occur). This framework focuses on the exploration of categorical spatio-temporal data (e.g., criminal incident reports where a type of crime, such as theft, is reported for a given location at a specific time). The goal is to enable the identification and exploration of hotspots and territories based on the underlying geographic network properties. To facilitate this, this framework, shown in Figure 5.2, consists of a central map view and several interactive widgets. When data is loaded into the system, categories are extracted and binned together over a user-defined time range and then assigned a primary color classification for rendering on the map. Analysts can interactively explore categories

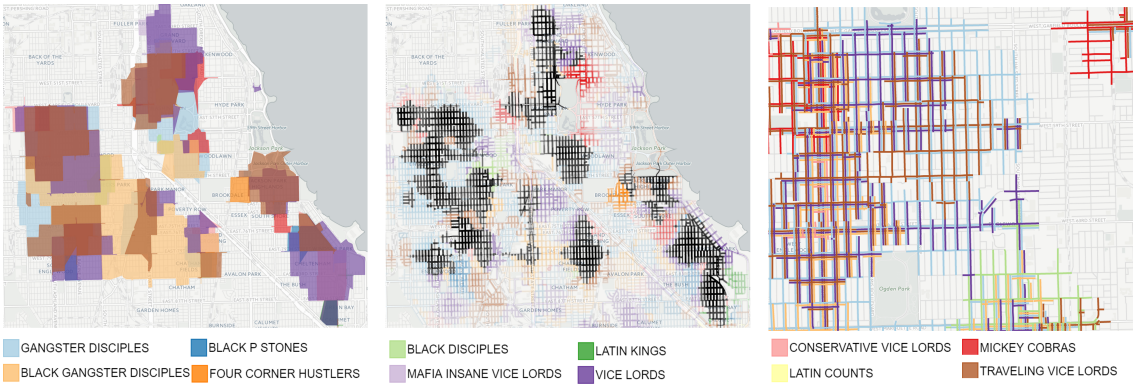


Figure 5.1: Gang Arrest Records in Chicago From 2014.

by toggling the color legend next to the category names, and the action buttons provide controls for exploring hotspots and territories. In the visualizations of this chapter, all KDE values are normalized into the range of [0, 1].

5.1 Data Description

Two kinds of crime incidents are used as test cases for the methodology. The first dataset is the crime records in Tippecanoe County, IN. This dataset is provided by the Visual Analytics Law Enforcement Toolkit (VALET) [111]. The second dataset is gang member arrest records provided by the Chicago Police. The location information includes street addresses and coordinates. The coordinates were generated with geocoding. In the second dataset, the location descriptions can also be the intersection between two roads. As the locations are reported by police input, the report might be approximate locations of the real locations. To project incidents to the networks, the points are firstly mapped the the projecting point on the nearest road segment. If the location is too far away from this segment, the distance is calculated with the lowest speed in the road network. In some cases, the geocoding result is based on the centroid of the building, so the distance might be inaccurate in the projection.

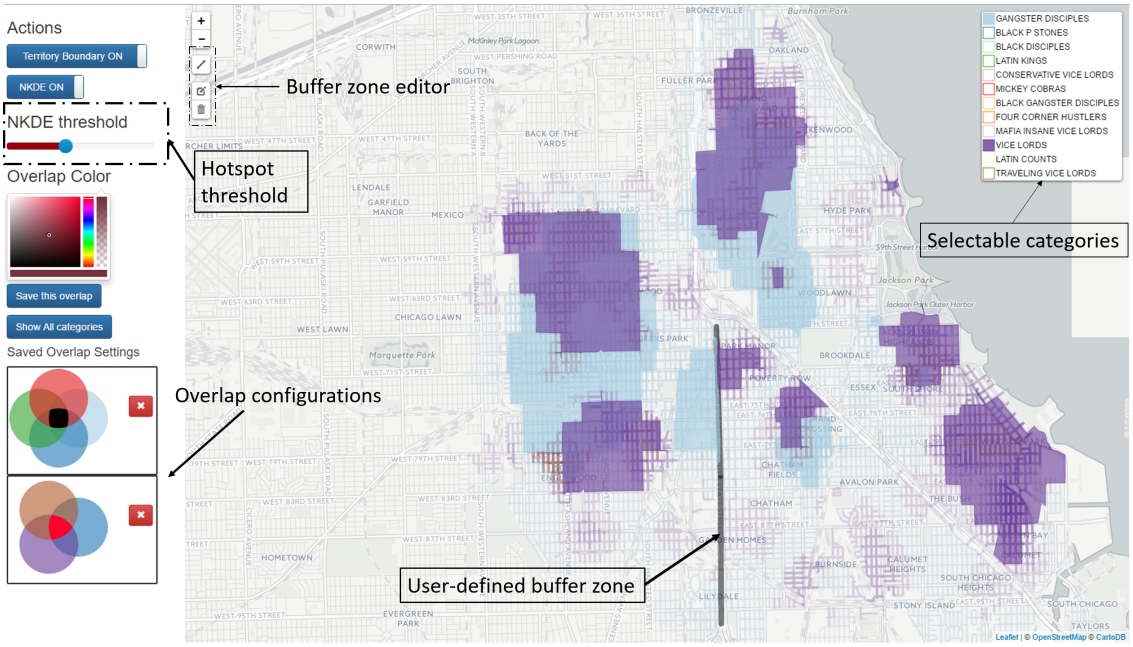


Figure 5.2: A Visual Analytics Framework for Hotspot Analysis Using Geographic Network Features.

5.2 Hotspot Visualization

One of the most common methods of visualizing hotspots is through the application of KDE to produce a continuous probability estimate from a set of sample points. The most common KDE implementation involves a fixed bandwidth estimation using a Euclidean distance function that approximates data across a 2D grid space. This implementation is seen in a variety of visualization tools and systems [106, 200, 201]; however, the visualization community has paid little attention to network based formulations of KDE. In order to situate our contribution in the literature, this thesis provide a brief overview of kernel density estimation, extensions of kernel density estimates to network geography, and then our formulation of edge weighted network kernel density estimation (NKDE).

5.2.1 Kernel Density Estimation

The general form of a kernel density estimator in a 2D space is:

$$\hat{f}_h(x) = \frac{1}{nh^2} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{5.1}$$

where $K(u)$ is the kernel function, n is the total number of samples and h is the bandwidth. A typical implementation involves gridding a planar area into small regions and counting the number of incidents per grid cell. Then, for each cell with a non-zero count, a kernel function, $K(u)$, is applied. This function acts as a distance decay effect for all sample points which spreads out the density to all cells within the bandwidth h. These values are then summed across all grid cells after the kernel function is applied to all non-zero count cells. The result is that the further the distance from a sample point, the less density will be accumulated.

A number of forms of kernel functions can be used to measure this effect, the most common being the Gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{5.2}$$

For simplicity, the Epanechnikov function (which is an approximation of the Gaussian kernel) is used in our experiments:

$$K(u) = \frac{3}{4}(1 - u^2) 1_{(||u|| \leq 1)}. \tag{5.3}$$

There is a large body of literature examining the impact of both the kernel choice and bandwidth selection on the resultant density estimation [104]. Results show that the choice of kernel has little impact on the resulting estimate when compared with the choice of bandwidth parameter, h. This parameter determines the smoothness of the estimated density (larger values of h give smoother estimations); however, too large of a bandwidth selection results in over-smoothing and too small of a bandwidth results in a very noisy estimate. Although bandwidth selection can be performed automatically [202], domain knowledge is

still very helpful in data analysis. Thus, the way in which a human would interpret distance must be codified by the bandwidth in the system. An intuitive distance measure would be the Euclidean distance, but, Euclidean distance is not reflective of true distance in many environments.

5.2.2 Network Kernel Density Estimation

To introduce kernel density estimation into the underlying geographic network, we need to estimate the density of a network-constrained space which can be formulated as a 1D KDE as shown in Xie et al. [115]:

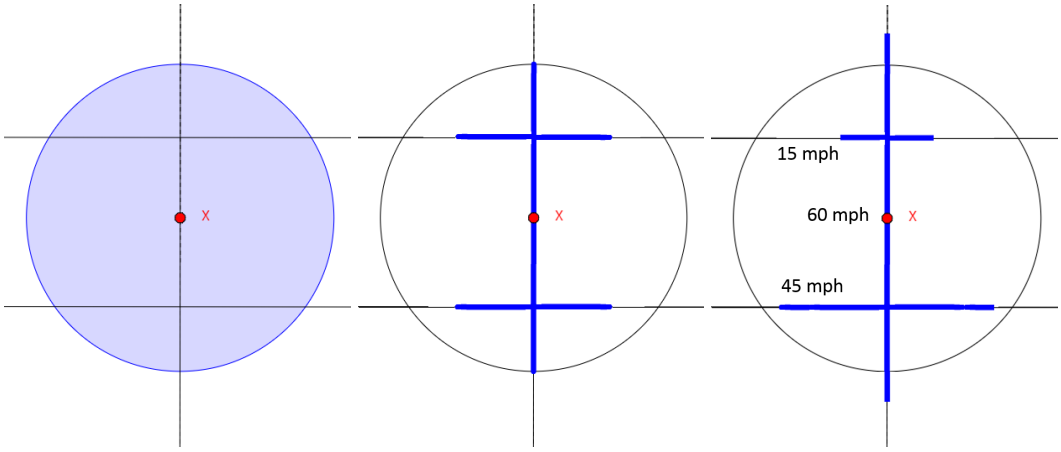
$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{N_x - N_{x_i}}{h}). \tag{5.4}$$

where N_x and N_{x_i} are the positions of lixel x and x_i in the network topology respectively.

Instead of calculating the density as a unit per area, it is now calculated as density over a linear unit and, like the pixel discretization schemes used in a 2D kernel density estimate, the network is discretized into equal-distance segments. Without loss of generality, these segments are considered to be unit length and the graph is undirected. For each sample point, x_i, the network is traversed from x_i until a distance h is reached. The density for each network on the segment is then calculated using the kernel function and the ratio of the distance of the segment x to x_i and the bandwidth h. The final density estimate is the sum of all kernel functions applied to all n sample points. For simplicity, it is assumed that all of the sample points x_i are mapped to nodes N_{x_i} in the network. Therefore, the distance between N_{x_i} and N_x is the length of the shortest path.

5.2.3 Modifiable Edge Bandwidth

The network KDE (NKDE) provides us with a density estimate based on the physical distance with respect to the underlying geographic network. However, the utilization of a



(a) Spatial KDE (b) Network KDE (c) Network KDE with Variable Bandwidth

Figure 5.3: Illustration of the Conceptual Differences Between the Density Estimation Algorithms.

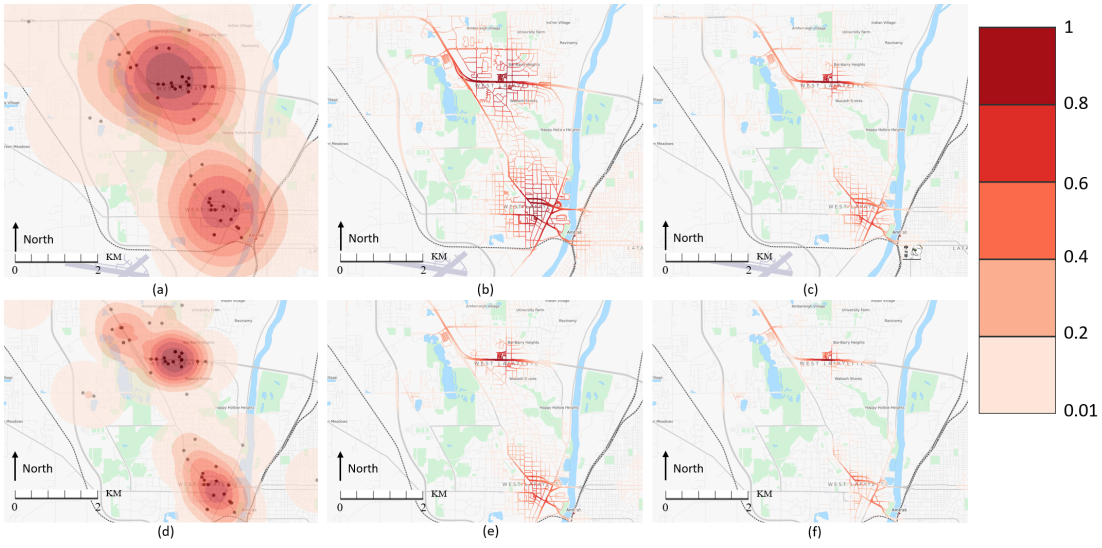


Figure 5.4: Traffic Accidents in West Lafayette, IN During March, 2014.

fixed bandwidth is still problematic. A fixed bandwidth on the network would only slightly better capture network properties than simply projecting the 2D kernel estimate values onto the underlying road network. In Figure 5.3 (a) it can be observed that the kernel for a single point would influence all the geography within a given radius. In Figure 5.3 (b), it can be observed that by using a network, not all regions on the network within the Euclidean radius will be influenced by a single point as distance is now measured along the network path. In Figure 5.3 (c), it can be observed that this effect is emphasized even more when edge effects are accounted for.

For this hotspot visualization approach, a formulation of NKDE to incorporate a variable edge bandwidth is developed. This formulation is similar to that of work by Downs and Horner [203] who use time geographic distance estimates to create probability densities, as well as work by Yu et al. [204] which modified KDE constraints for multi-factor networks. To introduce edge properties, such as the speed limit, we modify the bandwidth to be a weighted function based on the edge properties. As noted earlier, the kernel function is essentially modeling distance decay for the density spread. If edge weights are considered as impedance, the distance decay will be slower for some edges and faster for others. This leads to a bandwidth that is related to the edge properties. Now a fixed bandwidth h can be defined as the user controlled distance. This is the maximum distance decay that is allowed for a kernel. Then, the distance between nodes N_{x_i} and N_{x_j} can be defined as the length of the shortest path; however, in Equation 5.4, edges of the graph were unit length. In this formula, the edge length should relate to some physical property of the graph and a scaling factor such that the weighted edge length $we_{(i,j)}$ is defined as:

$$we_{i,j} = \frac{W}{G(e_{i,j})}|e_{i,j}| \tag{5.5}$$

where W is a user defined scale factor, and $G(e_{i,j})$ is the edge weight of $e_{i,j}$ from network G.

Next, the modified bandwidth d_i is introduced. First, the weighted distance of our sample points to all other nodes in the graph is calculated. Then, the longest unweighted distance whose weighted distance is less than or equal to h is picked. This unweighted distance is the modified bandwidth d_i. Thus, for sample point, i, the bandwidth changes based on the underlying network impedance giving us:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{K(\frac{x-x_i}{d_i})}{d_i}. \tag{5.6}$$

In the examples included in this thesis, the speed limit of an edge is used as the weighting property. Thus, edges with the fastest speed limit will have smaller lengths than those with lower speed limits. Several different weighting schemes can be used for the function, W, setting it to the maximum speed limit in the graph:

$$W = \max_{i,j}(G(e_{i,j})), \tag{5.7}$$

or to the average speed limit in the graph:

$$W = \operatorname{avg}_{i,j}(G(e_{i,j})), \tag{5.8}$$

where $G(e_{i,j})$ is the speed of $e_{i,j}$. However, W could be interactively chosen based on domain knowledge of the user. Such a formulation is directly applicable to the visualization community as geographically rich datasets are becoming more abundant, and modeling hotspots using the appropriate underlying geographical features is becoming more and more critical for advanced analysis.

Figure 5.4 illustrates the differences between the three density estimators using real-world traffic incident data to underscore the need to appropriate visual representations. The bandwidth in Figure (a, b, d, e) is 1 mile and the bandwidth in (c, f) is weighted by the edge speed limits such that a point on a road with a 60 mph speed limit will spread 1 mile, or $W = 60$ in Equation 5.5. Let us compare our modified NKDE method, Figure 5.4

(c) with the fixed bandwidth NKDE of Xie et al. [115] shown in Figure 5.4 (b). Note the local variations that are apparent in (c) that are not apparent in (b). For example, the areas noted as University Farm and Bar-Barry Heights in the northeast corner of the map have significantly reduced hotspots when estimated with our variable edge bandwidth, and the on-ramps to highways in the southern hotspot (the rounded curves near the T in West Lafayette) have a significantly reduced density as well. These local variations seem to better match the expected density values as fewer accidents are expected in low speed neighborhoods and on clover-leaf exchanges. These local variations are even less apparent in the most commonly implemented KDE method, Figure 5.4 (a). While both Figures 5.4 (b) and (c) map the KDE to the underlying geographical network by taking into account the edge properties, more local variations may become apparent as shown in our formulation of NKDE.

With the instantiation of NKDE in our framework, users can further explore hotspots with interactive level sets. In Figure 5.2, users can adjust the hotspot threshold slider, and road segments with NKDE values lower than the chosen threshold will be rendered as normal. In this way, users can also explore the impact of local network structures on the spread. As the NKDE threshold increases or decreases, the hotspot visualization will change based on underlying network edge properties. In the case of the example images, the effect is based on posted road speed limits.

5.3 Territory Extraction and Overlap Analysis

By utilizing network properties, the proposed framework enables analysts to identify hotspots related to network features that may play a role in the underlying geographic phenomena. Furthermore, NKDE can also be applied to extract territory boundaries. When speaking to police officials, one common way of describing areas of criminal activity is by defining regions by road boundaries; for example, north of Interstate 290, south of Inter-

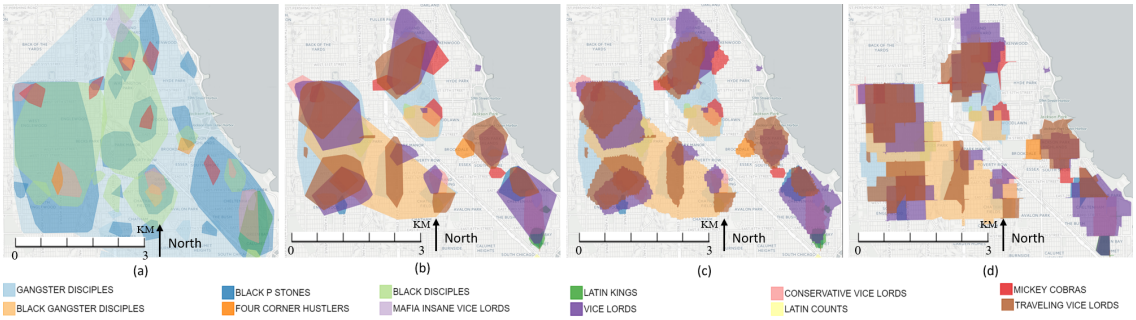


Figure 5.5: Comparison Between Results of Different Stages in the Territory Extraction Algorithm.

state 90, or east of Highway 43 is a common way of describing territory boundaries. Thus, using simple convex hulls or merely using threshold extraction from traditional KDE will result in boundaries that do not necessarily follow roads. Furthermore, research has also shown that an analysis of territories is useful for analyzing various types of crime. Huddleston et al. proposes a multi-level territory estimation algorithm [205] for identifying gang spheres of influence. Work by Nakamura et al. [206] and the Rand Cooperation [207] shows that competition between gangs can potentially lead to more violent conflicts. As such, different kinds of crimes can also have spatial correlations which can be shown with territory overlap. Thus, this framework utilizes a novel territory boundary extraction algorithm for the analysis of the correlation and interactions between categorical spatial data (i.e., categories of criminal activity such as robbery and public intoxication or categories of gang members).

5.3.1 Territory Extraction

Given a set of geolocated points, the most straightforward method to define and extract territory boundaries is to simply draw a convex hull around the points. However, this method often overestimates or underestimates the area with results being sensitive to outliers. Furthermore, a naive convex hull solution misses the fact that a territory may not

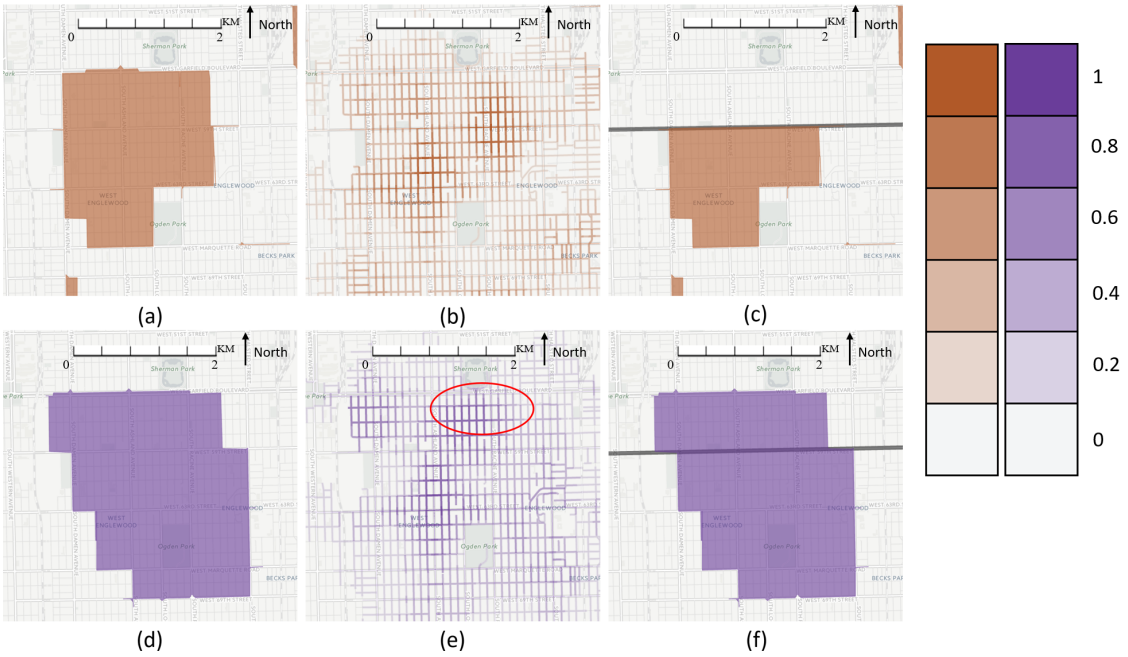


Figure 5.6: Interactively Editing Buffer Zones Can Reveal Underlying Structures in the Territories.

be contiguous, i.e., a group might have several small territories scattered at different locations, which will result in a gross overestimation of the territory. As such, more robust methods have been developed to extract spatial clusters (territories/hotspots), with one of the most popular being DBSCAN [208]. However, the convex hull around points within one cluster of the regions found with DBSCAN will focus on the individual locations of the phenomena, as shown in Figure 5.5 (a). This neglects the potential spread of territories.

More recent work has demonstrated that level sets applied with kernel density estimation provide reasonable estimates for territory extraction as well (e.g. [209–211]). Unfortunately, as previously discussed, traditional KDE does not take into account many of the physical geographic properties that may be critical for analysis. Thus, we propose a method for territory extraction using the underlying spatial network structure that combines elements of NKDE and DBSCAN.

First, NKDE is performed, and each point in the dataset is assigned a density estimate value. If the NKDE value of the point is greater than a threshold, the point is extracted and clustered with DBSCAN. Next, convex hulls are extracted around the resulting clusters. The convex hulls are shown in Figure 5.5 (b). Finally, the edges of the convex hulls are discretized and mapped to the geographic network.

In order to map the convex hulls to the geographic network, a force directed method is applied. In the physics model, each point in the DBSCAN cluster has a positive point charge with quantity q_1 and each network edge is a rod with a continuous negative charge. Thus, all points will be attracted to the network edges. Next, the edges are simplified as negative point charge with a quantity q_2. Then, the force of attraction between points and network edges is calculated with Coulomb's Law:

$$F \sim \frac{|q_1 q_2|}{r^2}, \tag{5.9}$$

where r is the distance between the point and street. When the forces from the nearby streets are calculated, a winner-takes-all strategy is applied and the point is moved to the street which contributes the largest force. The quantity of charges on the edges can be determined by a variety of design factors. If the roads are charged uniformly, the results are shown in Figure 5.5 (c). For this work, information on the level of the roads is used (thus the convex hull boundaries will snap to larger roads first, i.e. highways). We also provide interactions so that users can mark roads as critical, thus adding a larger charge value to the marked roads. The results are shown in Figure 5.5 (d).

As with many territory extraction algorithms using density estimates, the hardest part is determining an appropriate threshold value. Our framework supports interactive threshold modification through the NKDE threshold slider where the user can explore different hotspot thresholds to find an appropriate upper and lower limit based on:

$$y_i = (u-l)(1+\log_{10} min(x) - \frac{\log_{10} x_i}{\log_{10} \frac{max(x)}{min(x)}}) + l, \tag{5.10}$$

where y_i is the threshold of group i with size of x_i, and u and l are the threshold upper and lower limits. When the territory boundaries are enabled in the framework, changing the threshold will directly impact the size of the territory that is extracted. The analyst can use their domain knowledge about the region coupled with the interactive threshold widget to identify values that best match their mental model of the area.

5.3.2 Overlap Analysis

The proposed territory extraction method provides the spatial regions for a particular category of the data, e.g., the regions where thefts are most likely to occur, or the regions where the *Black Disciples* are most commonly active. These regions can then be used to quantify the amount of overlap between the different categories found in the dataset, (e.g., how often do public intoxication and noise complaints overlap?) Since these regions contain a probability distribution estimate, we can utilize the Bhattacharyya Coefficient *BC* which is defined as

$$BC(p_1, \ldots, p_n) = \int \sqrt[n]{\prod_{i=1}^{n} p_i(x)} dx, \tag{5.11}$$

where $p_i(x)$ is the probability density function of ith distributions [212, 213]. If no overlap settings are saved and all categories are selected, the pair of categories with the maximum *BC* will be used to encode the overlap intensity. Otherwise, the n-distribution Bhattacharyya Coefficient is visualized.

As shown in Figure 5.1, this system creates a map that shows multiple categories of data as well as encodes regions of overlap. The left figure shows estimated gang territory boundaries with overlapping territories highlighted. The middle figure shows a geographic

based network view of gang arrest records using NKDE and overlaps are again highlighted. The right figure shows a detailed view of the overlap areas. However, it can be difficult to determine exactly which categories are overlapping, thus, this system also provides a detail zoom view (Figure 5.1 - Right) where the corresponding color of each overlapping category is now mapped on the network segment. The overlapping groups are ordered from top to bottom or left to right (depending on the network orientation) with respect to their density in a region, for example, in Figure 5.1, the streets within the territories are mapped to the corresponding colors. If there is only one category (e.g., gang in Figure 5.1) on that network segment, then only one segment is drawn along the street. If multiple categories overlap in that region, multiple segments are aligned by the streets. The segments are ordered by the category densities and listed left to right or upwards.

One issue is how to appropriately visualize the hotspot overlaps on the map. Since the data is categorical in nature, this overlap can be viewed as a set problem. Our framework provides an interactive overlap configuration widget, as shown in Figure 5.2. First, the user selects the categories of interest and chooses a color to define the overlap. The user then selects "save the overlap" and the set intersection is visualized below. When this configuration is selected, the map will re-render such that the overlap coefficient is defined based on the categories related to the selection. The opacity of the overlap color is then directly proportional to the Bhattacharyya Coefficient.

Unfortunately, the combined NKDE - DBSCAN territory extraction method is unable to account for potential natural geographic boundaries in the data. In such a case, details as to which geographical features break the network flow would need to be encoded, and such information is generally only captured in the domain knowledge of analysts. For example, police may know that certain gangs do not operate north of a particular highway or railroad. In order to incorporate such knowledge into our hotspot analysis and territory extraction methods, the user is allowed to define buffer zones on the map using our buffer

zone editor tools, shown in Figure 5.2. These buffer zones serve as cuts to the underlying geographic network, which then limits the directionality of NKDE and directly influences the boundaries of the extracted territories. Specifically, if the user draws a cut and the territory/hotspot becomes smaller, then the reduced part was not much of contributor to the underlying NKDE value. This indicates that the part of the territory that disappeared is likely not an important contributor to the overall territory shape. Thus, while events may occur outside of this territory, the major boundaries can be defined in this manner. Figure 5.6 illustrates the addition of two edge cuts in the map. In Figure 5.6 (a-c), the cut is added and the territory shrinks, indicating that the edge is a reasonable boundary for the region. This is due to the fact that the NKDE values can no longer reach beyond the cut, so sparse regions will have their overall density values reduced. This results in those regions no longer being identified as hotspots, thus they are removed from the hotspot territory extraction. In Figure 5.6 (d-f), the cut is added and the territory remains the same, indicating that the underlying point distribution is large on both sides of the network cut. As such, the buffer zone that was indicated is unlikely to actually exist.

5.4 Case Studies

In order to evaluate our proposed framework, two datasets focusing on criminal incident reports in Tippecanoe County, IN and gang violence in Chicago, IL are used. For our edge properties, speed limits are utilized. Previous research [214] indicates that crime (such as burglary) is related to the accessibility of a neighborhood and part of the measure of accessibility is the distance of travel. Colquhoun found that crime rates were lower in areas with complex road networks and lower designated speed limits [215]. As such, the application of road speed for the weights in our modifiable edge bandwidth NKDE should be a reasonable metric to help capture physical barriers (as suggested by Rossmo [216]).

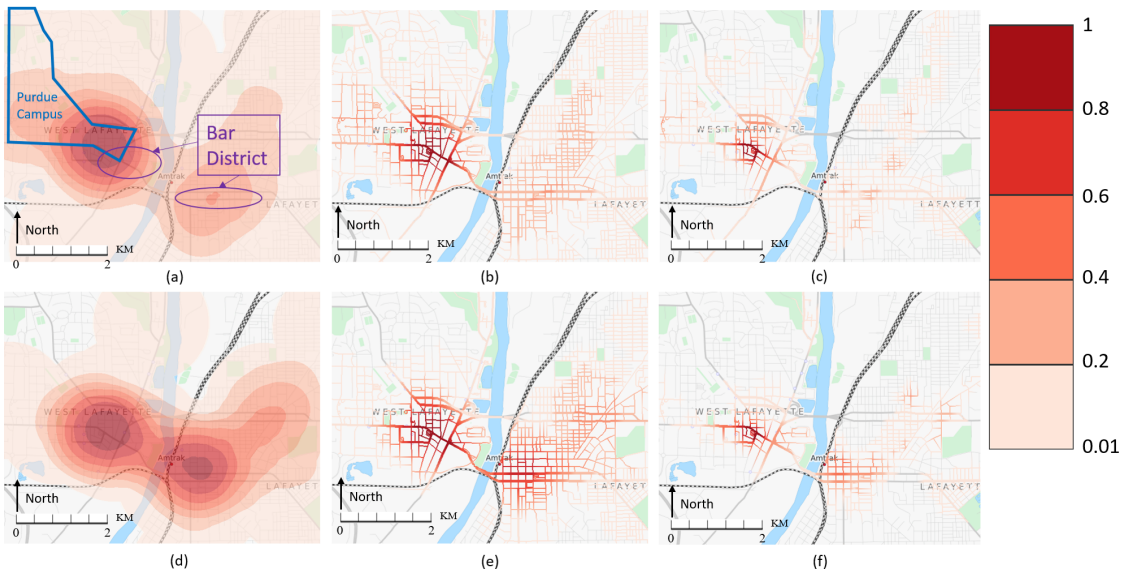


Figure 5.7: KDEs for Public Intoxication Incidents in Tippecanoe County.

5.4.1 Criminal Incident Reports and NKDE

In this case study, the user explores differences in the geographic distributions of criminal incident reports in Tippecanoe County, Indiana, the home of Purdue University. A total of 21,332 reports comparing the Spring and Summer of 2014 (January 10 – May 10 and May 11 – August 20 respectively) are used. For this study, a series of images are created using the traditional kernel density estimation approach as well as the variable edge bandwidth network kernel density estimation approach. Results were presented to a Purdue University police officer, and he provided an interpretation of the results and commented on how he would use the information. For the edge weighting, the speed limits of the roads are used for the scaling factor. The average speed limit in the area under analysis is 30mph, thus $W = 30$ in these examples.

In Figure 5.7, public intoxication incidents and identified landmarks are visualized for the discussion. Figure 5.7(a-c) visualizes the incidents in spring of 2014 and Figure 5.7(d-f) shows the summer of 2014. The fixed bandwidth is 0.5 miles and the variable bandwidth

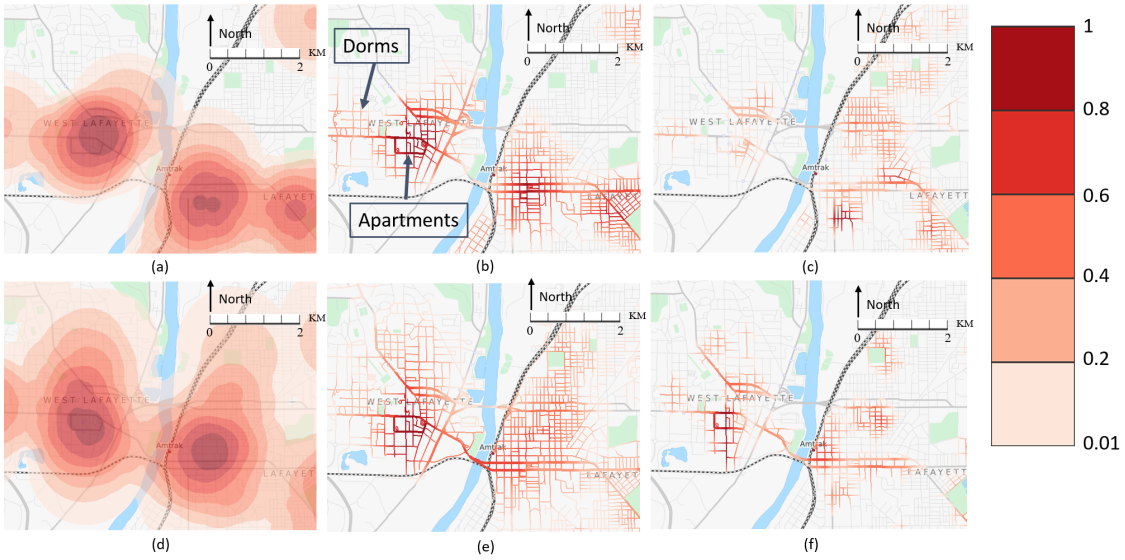


Figure 5.8: KDEs for Bike Thefts in Tippecanoe County.

is $W = 30$ corresponding to the average road speeds in the major hotspot areas. From the traditional kernel density estimation pictures, the officer described his impressions of the two hotspots separated by the river. The western hotspot is located near the campus downtown region and the traditional bar district for students. The eastern hotspot is also a bar district; however, the distance from campus typically requires driving to this area. The officer noted that the distinction between the spring hotspot in the east being smaller than the summer was likely due to the large amounts of festivals held on the eastern side of the river during the summer months. After that, the officer looked at the network kernel density estimation maps Figure 5.7 (b) and (c). He commented that the immediate benefit for this was in directing targeted law enforcement. Looking at the maps of the communities of public intoxication on the west and east sides of the river, the officer noted that these paths map to common parking structures and undergraduate dorms in the data set. By doing random foot patrols optimized by these locations, officers can potentially prevent drunk driving. For the variable bandwidth shown in Figure 5.7(c,f), notice the subtle local variations where the density does not cover all of the neighborhoods in the eastern part of

the town and instead it follows primarily along the main travel corridors as compared to Figure 5.7(b,e). Such information seems to better conform to the officer's mental model of public intoxication arrests being more localized. Again, this does not imply that one density estimate is superior to another. But by adding in more information to the model, local variations can be explored.

Next, the officer explored the results of the bike thefts in the spring as shown in Figure 5.8. Figure (a-c) visualizes the incidents in spring of 2014 and (d-f) shows the summer of 2014. The fixed bandwidth is 0.5 miles and the variable bandwidth is $W = 30$ corresponding to the average road speeds in the major hotspot areas. While the hotspot of Figure 5.8 (d) appears to cover all of the campus in the west, the road network actually enables the officer to tell a different story based on his experience. In the bike thefts near campus, the area of town where upperclassmen are likely to live (apartments) and the area where underclassmen are likely to live (dorms) are marked. The visualizations show that bike thefts are taking place primarily in the upperclassmen area. He also noted that this network view provides a nice way to communicate public awareness campaigns as well as patrols. Specifically, he noted that there were very few bike thefts in the north of campus and could communicate that better with the network based images than with the traditional hotspot map. Again, note the subtle local variations in the fixed bandwidth network KDE and the proposed variable bandwidth method. Here the fixed bandwidth network estimate spreads the data across all neighborhoods similar to that of the Euclidean version, whereas the proposed method takes into account edge properties and gives potential insight into local variations.

Overall, the feedback was positive, and, from the examples discussed, it was clear that the network view could provide insights into geographic hotspots that would not be obvious from the purely Euclidean based methods. As the bandwidth shrinks, the visualization approximates that of a point based map, and so even if one were to change the bandwidth

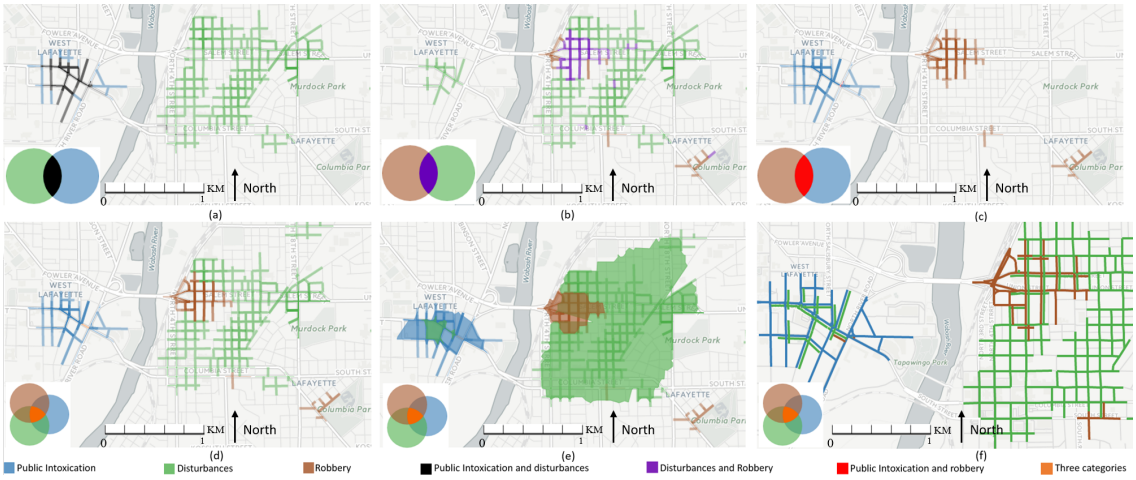


Figure 5.9: Overlap Analysis of Public Intoxication, Disturbances, and Robberies in Tippecanoe County

parameters of the Euclidean distance estimation, it would also fail to show distributions along the road. As such, there is a need for analytical techniques that can combine multiple properties of complex spatial data into a single view.

5.4.2 Spatial Overlaps in Crime Categories

In this example, the analyst is exploring criminal incident reports in Tippecanoe County categorized into theft, public intoxication, disturbances, burglary, noise, and robbery. The analyst is first interested in understanding the different crime hotspots. The analyst cycles through the different crime categories to gain an overview of where different crimes take place, reconfirming their mental model and identifying major roads of interest. After adjusting the NKDE threshold, the analyst is satisfied that the resulting hotspots are of interest and extracts all three territories on the map view (Figure 5.9). Next, the analyst wants to explore set overlaps between the different crime types. The initial impression is that the overlap will be primarily two independent sets, with public intoxication and disturbances on the west side of the river and disturbances and robbery occurring on the east side of

the river. The analyst switches the territory boundary off to observe the joint probability metric between the different categories. First, the analyst looks at the set intersection between disturbances and public intoxication and chooses the overlap color to be black, as illustrated in Figure 5.9 (a). This intersection occurs in close proximity with the known bar districts in town, indicating that such infrastructure may directly correlate to these types of events. Next, the analyst explores the intersection between disturbances and robbery. The analyst defines another overlap color, purple, as illustrated in Figure 5.9 (b), and finds that the area of intersection between disturbances and robberies are in a completely different part of town than disturbances and public intoxication. This area of overlap occurs in the lower income portion of town. Next, the analyst explores the intersection between public intoxication, and robbery as illustrated in Figure 5.9 (c) and defines the overlap color to be red. Little overlap between public intoxication and robbery are found to occur using the current NKDE threshold values. Finally, the analyst explores the intersection of disturbances, public intoxication and robbery simultaneously, creating a fourth overlap color of orange as illustrated in Figure 5.9 (d). The analyst then zooms into the different regions looking at streets where the major overlaps occur as illustrated in Figure 5.9 (f). The analyst observes that the primary region where all three types of crime intersect are in the low income housing area, but again, the overlap opacity is very low indicating a low probability of overlap. However, the analyst does observe an area of orange overlap on the west side of the river that was not apparent in looking at the territory view alone. The analyst zooms into the location and observes a single block that appears to be the main intersection of these three types of crimes.

The main benefit of the overlap coefficient is that the probability metric is not solely dependent on where the spatial territories overlap. For example, Figure 5.9 (e) shows the convex hulls of the territories for public intoxication, robbery, and disturbances. If one were to only look at the convex hulls, it would appear that the overlap between public intoxica-

tion and disturbances are completely independent from public intoxication and robbery. However, the territory extraction is a density based scan coupled with NKDE metrics. By visualizing NKDE and overlap coefficient measures, the underlying data distributions can be further explored. What is most interesting about this type of analysis is the ability to quickly explore correlations between locations. While previous work [217] has explored temporal correlations between locations to look at lag and lead aspects of crime spreads, this method provides a new tool for analysts where they can search for spatial correlations in the form of set probabilities. Future work will explore adding a temporal dimension into the overlap calculation to be able to suggest lag and lead. Further future work will explore extracting point-of-interest information as a way to annotate locations in the data to help further develop mental models of hot spot territories. For example, analysts may want to know how many bars are in a given territory or how many bus stops are within a given proximity.

5.4.3 Gang Territory Analysis

While the previous examples demonstrate how different types of crime can be explored for spatial correlations, another major use of territory analysis is for gang identification. When discussing with police, one of the main methods for creating gang territory maps is by plotting arrests and graffiti and the drawing boundaries by hand. In this example, a data driven approach is used. The gang territories are plotted with the proposed method and then some boundary cuts along major geographic features (interstates, light rail, etc.) are added interactively. Figure 5.10 shows the interactive process to extract territories of Chicago Gangs. The top row explores buffer zones and their impact on defining gang territory edges on Interstate 94. The bottom row explores Washington Park as a nexus for gang activity, using the park itself as a buffer zone in exploring territory overlap. In this analysis, we choose to focus on overlaps between the *Gangster Disciples*, *Vice Lords*, and *Black*

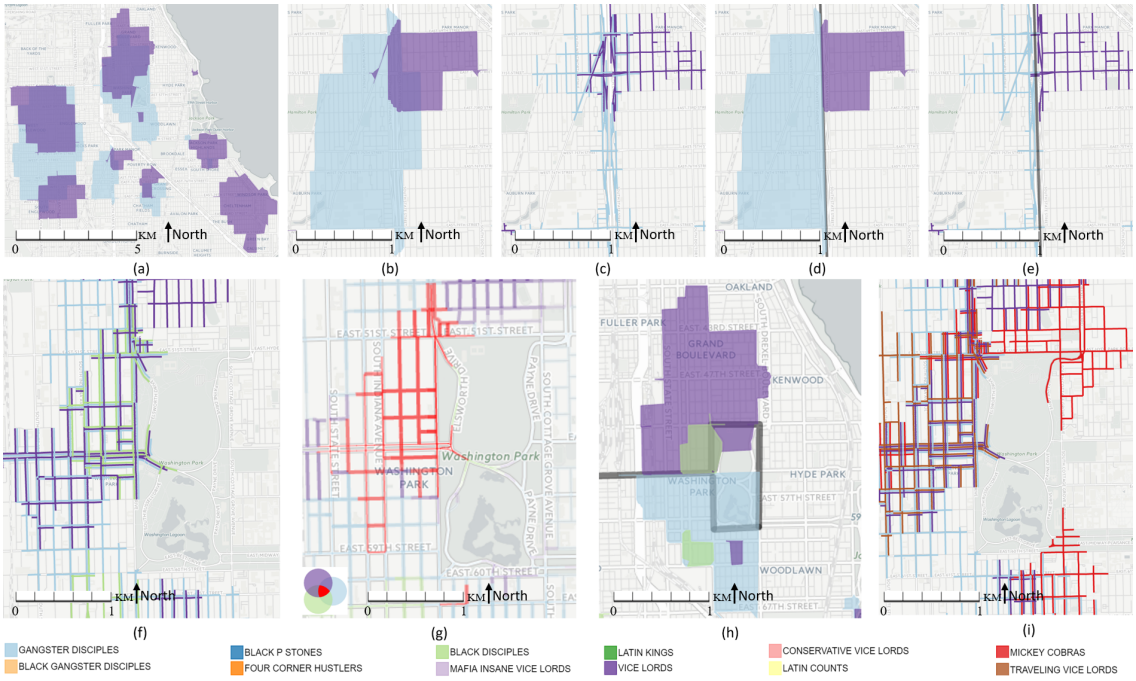


Figure 5.10: Territory Overlap Analysis of Rival Gangs in the City of Chicago

Disciples as these gangs represent the three largest gangs in Chicago. First, the overlaps between the *Gangster Disciples* and the *Vice Lords*, Figure 5.10 (a), are explored. Here, large areas of overlap between the two gangs are visible; however, there are places where the overlap seems tangential to nearby infrastructure, specifically parks and interstates. As such, the user can interactively explore the territories by drawing buffer zones along several major roadways. It can be seen that when a buffer zone is drawn near the Park Manor area, the territories are split and so it appears that Interstate 94 may serve as a natural boundary for the gangs as shown in Figure 5.10 (b-e).

Next, we explore the overlap between the *Gangster Disciples*, *Vice Lords*, and *Black Disciples*. From Figure 5.10 (f-g), it can be seen that the Washington Park area is serving as a major intersection area between the three gangs. We add an edge cut around the park and can find that there is some separation between the gangs as shown in Figure 5.10 (h); however, Washington Park is at the nexus of the territories. While the edge cuts serve to

highlight major boundaries and we see relatively strong separation between gang communities, further investigation finds that Washington Park is ranked 5th among Chicago's 77 community areas with respect to violent crime in the last month (http://trib.in/2mLwSNI). To further explore problems in this area, the user can toggle other gang territories on and see that several other gangs also have territories proximal to Washington Park in Figure 5.10 (i).

What is interesting about this work is that by interactively adding and removing edge cuts, the combination of DBSCAN and NKDE provide a means of exploring how sensitive these territories are to different geographical features. One simple exploratory exercise is to simply mark several of the major secondary streets as edge cuts and explore the resulting territories and NKDE values. Similar to the crime type overlap in Tippecanoe county, the user can also create an overlap map and interactively define unique color classes for intersections between gangs. One could hypothesize that regions of overlap could serve as a nexus for violence. Anecdotally, one major rivalry of late was between the Black Disciples and Gangster Disciples in 2012. Aspiring rapper Lil Jojo was killed in a homicide which was attributed to gang violence through reports on social media [218]. The crime site locates in close proximity with the overlap regions between these two gangs and marked in Figure 5.11. Figure 5.11(a) shows that that these gangs share several regions where the territory boundaries overlap. In Figure 5.11(b), further inspection using NKDE shows they have a high probability of overlap along a large number of roads. Several high profile deaths have been linked to rivalries between these gangs in the ✸ region. This area is also where the actress Jennifer Hudson's family was murdered (that killing was also attributed to gang violence) [219].

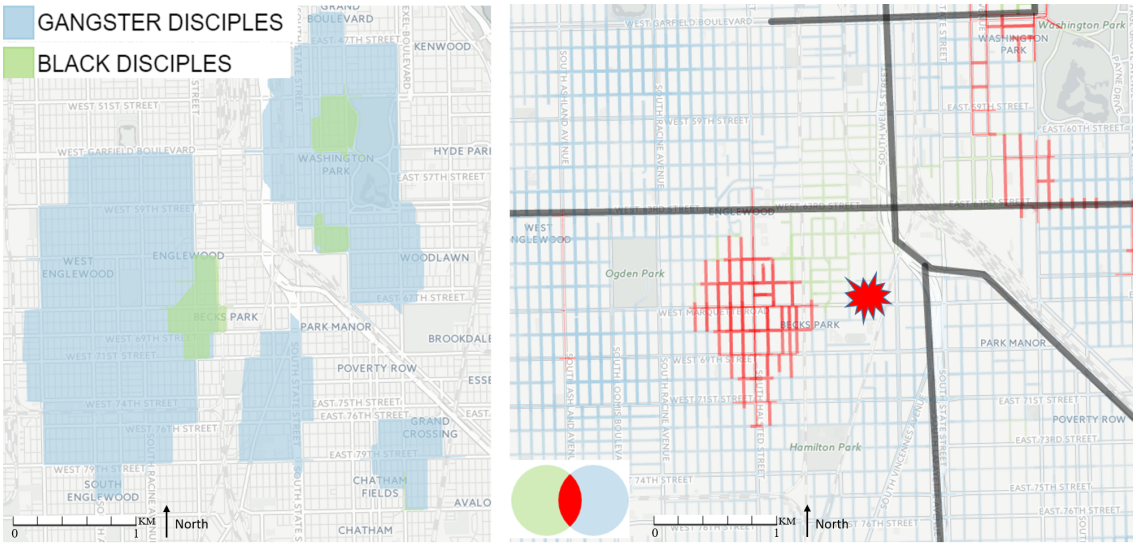


Figure 5.11: The *Black Disciples* and *Gangster Disciples* Territories and Their Overlaps

5.5 Evaluations

In recent studies of predictive policing, it has been more popular to analyze incidents and arrange patrols in the unit of street segments and intersections (e.g. [126, 127]). For hotspot analysis, we can get the hot street segments with either NKDE or spatial KDE projected to the road network. To evaluate the effectiveness of KDE in hotspot identification, two kinds of analysis are applied. First, spatial KDE and network KDE are compared in quantitative and visual approaches. As network KDE has higher computational complexities, the comparison can be used as guidance for when NKDE can be estimated with spatial KDE at a lower cost. Then, these two KDEs are evaluated using various metrics to measure the effectiveness and accuracy in spatial analysis. For fair comparisons, spatial KDE results are projected onto the road network.

5.6 Comparison of KDE

The first evaluation is the comparison between spatial KDE and network KDE. The comparisons are performed in both quantitative and visual aspects. First, quantitative differences between two KDEs are calculated under different settings. Because network KDE has higher computational complexity, it can be estimated or replaced by spatial KDE. If the difference is significant, visual comparison is performed to analyze the difference in detail.

5.6.1 Quantitative Comparison

To compare two KDE algorithms, the overall differences are first summarized with mean squared error (MSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{5.12}$$

where $\hat{Y}$ and Y are two image vectors. Each value in the image represents the KDE value on the corresponding pixel. The real value of KDE, rather than the rendered color, is used in this case because the focus in this stage is to evaluate the difference between values instead of the final render results.

As noted in the previous section, one of the issues of KDE in analysis is how to determine the bandwidth. As KDE models the dispersion procedure, the differences between NKDE and spatial KDE can unveil the patterns of the disperse. The first set of comparisons focus on the differences between different types of crimes in different crimes.

The first example is to visualize the differences between two KDEs generated from incidents in Tippecanoe County, IN. The RMSE differences are visualized in Figure 5.12. The horizontal axis is the spatial KDE bandwidths in miles. The vertical axis is NKDE bandwidths in seconds. Darker colors denote larger differences. Along the vertical axis of spatial bandwidths, it can be observed that the two KDEs reach more agreements on the lower half of the bandwidths. The bottom generally lies near the minimum speed limits

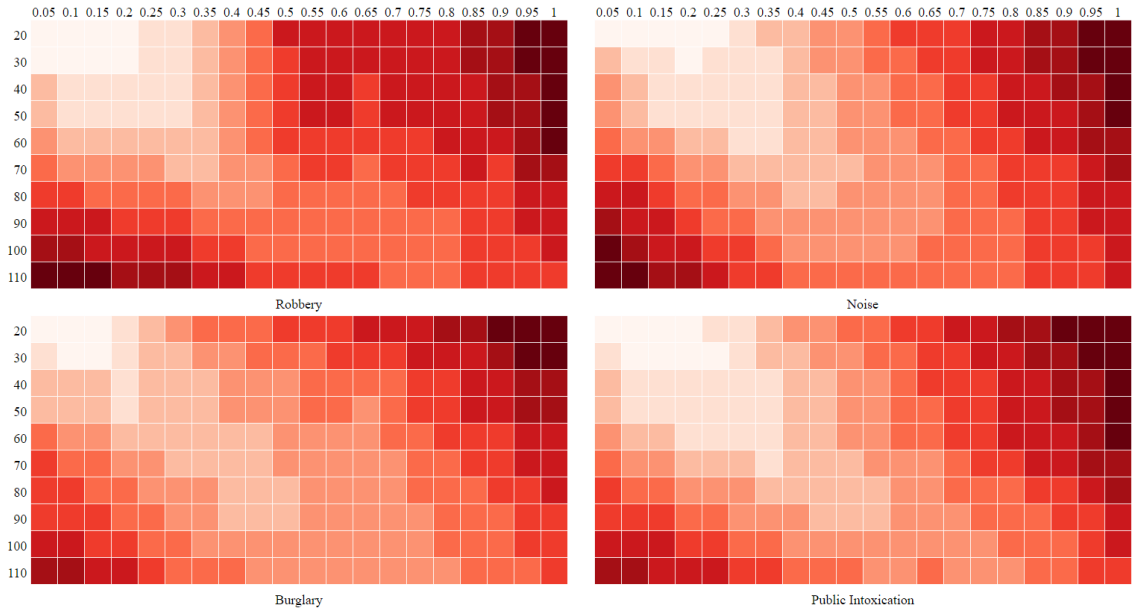


Figure 5.12: RMSE Differences Between Spatial KDE and NKDE in Different Bandwidth Settings for Four Types of Crimes in Tippecanoe County, IN

(0.2 miles for 30 seconds and 0.35 miles for 60 seconds). Note that spatial KDE models' dispersion in the straight line and the networked KDE disperses the incidents along the road. Another insight we can gain from this finding is that the crime incidents in this area might have less mobility than speed limits. This could be related to the nature of the incidents types. Many of the events are more likely to be caused by people on foot instead of in vehicles. As a comparison, the same analysis is applied to Chicago Gang arrest records as shown in Figure 5.13. The observed lowest point shifts to 0.4 miles for 60 seconds network bandwidth. As Chicago is a larger metro area and the gang members have higher mobility, it seems likely that more incidents of gang related crimes are mobilized with vehicles instead of on foot. This comparison shows that average speed limits can be a good starting point of the analysis for NKDE, but it should be adjusted according to the applications.

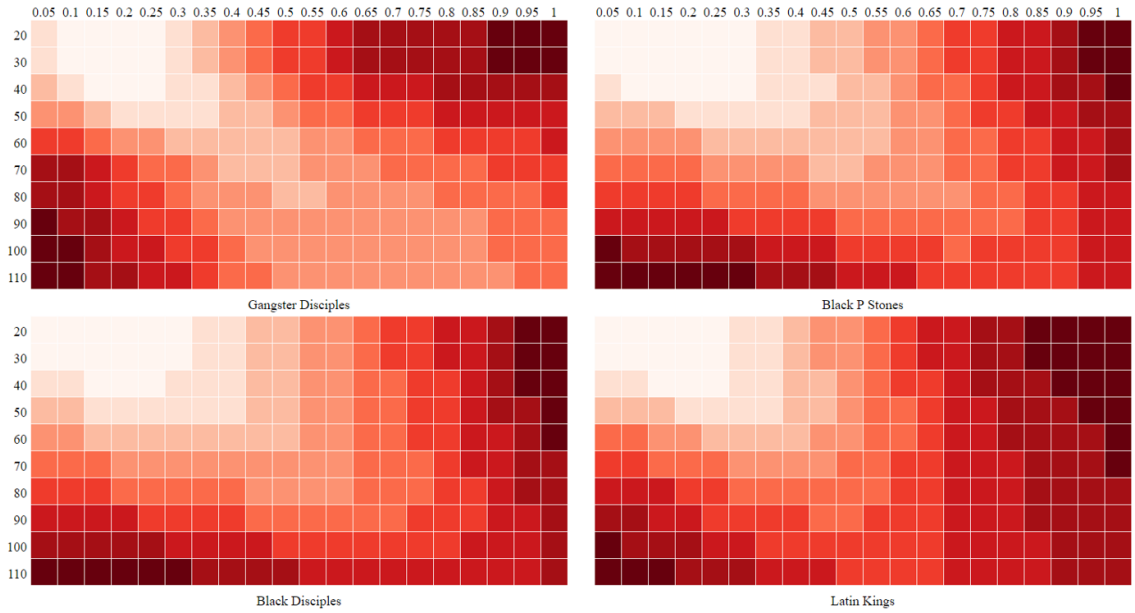


Figure 5.13: RMSE Differences Between Spatial KDE and NKDE in Different Bandwidth Settings for Arrest Records of Four Gangs in South Chicago.

5.6.2 Visual Comparison

With the quantitative comparisons, we can estimate the appropriate bandwidth values for spatial KDE and NKDE for fair comparisons. The next question is to determine which method can generate better hotspot visualizations. In this section, the differences of the KDEs are visualized with real data points overlay. For the incidents in Tippecanoe County, IN, the bandwidths for spatial KDE and NKDE are set as 0.2 miles and 30 seconds respectively. Red and blue colors denote larger values for NKDE and KDE, respectively. Figure 5.14-5.16 compare NKDEs and KDEs of three types of incidents in Tippecanoe County, IN. In Figure 5.14, disturbance incidents are used as the test case. The first row shows the overview of the KDEs. The left image is NKDE. The center image is KDE. The right image is the difference between NKDE and KDE. Red colors denote that NKDE has larger values and blue colors denote smaller values. The second row shows the zoomed views of the local details. The first two images show the areas where NKDE is larger. The

Figure 5.14: The Comparison Between NKDE and Projected Spatial KDE With the Disturbance Incidents.

third image shows the area where spatial KDE has larger values. The last image shows the area where two KDEs have similar values. Four detailed areas are zoomed in to show the difference. In the first two detail figures, concentrated roads get more attention from NKDE. In the third figure, a street gets more attention from KDE, but it can be seen that there are not incidents in the east side of the blue roads. Figure 5.15 compares the two KDEs with the noise incidents. The left image is NKDE. The center image is KDE. The right image is the difference between NKDE and KDE. Red colors denote that NKDE has larger values and blue colors denote smaller values. In the right figure, the main streets in the southeast are recognized as hotspots by spatial KDE. However, there are not many incidents around this area. Figure 5.16 compares the two KDEs with the case of burglary incidents. In the right figure, the spatial KDE gets better performance than the other two cases. However, it can be observed that the distribution is more even within one community and different across different areas. So the reason could be the difference of location selection schemes.

Figure 5.15: The Comparison Between NKDE and Projected Spatial KDE With the Noise Incidents.

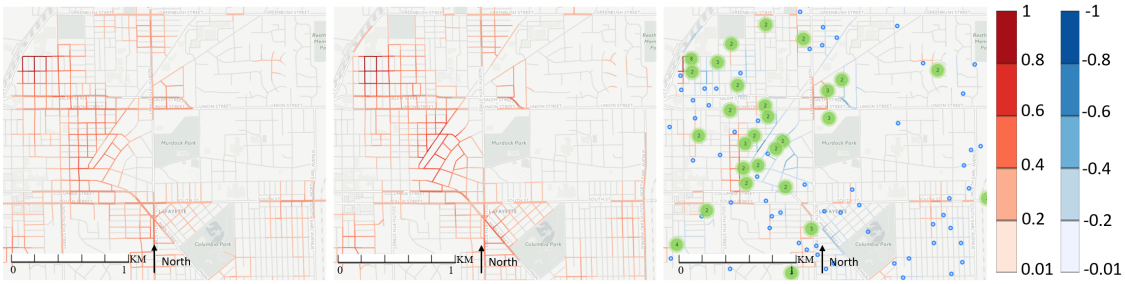


Figure 5.16: The Comparison Between NKDE and Projected Spatial KDE With Burglary Incidents.

5.7 Accuracy Evaluations

The accuracy metrics are based on the hotspot prediction tasks. One of the most popular evaluation methods is cross validation. Traditionally, cross validation is used to pick the optimal bandwidth values acoording to the stability of KDE [220–224]. They include Likelihood Cross Validation (LCV) [225, 226], Least Square Cross Validation (LSCV) [227–229], Biased Cross Validation (BCV) [230], Smoothed Cross Validation (SCV) [231], and the direct-plug-in method [232]. In cross validation, the error can be measured in various metrics such as integrated squared error (ISE) [233, 234] and L_∞ [235]. For example,

leave-one-out LSCV with the metric of ISE can be formulated as:

$$ISE = \int (\hat{f}(x) - f(x))^2 dx \tag{5.13}$$

where $\hat{f}(x)$ is the density function for position x in the investigated dataset and $f(x)$ is the density function in the overall investegated [234]. As leave-one-out cross validation is used, $\hat{f}_{-i}$ can be denoted as the density function for the dataset generated by removing observation X_i. LSCV is the cross validation process to minimize average ISE for all validations. LSCV is the average of ISE for all n cross validations.

In the accuracy evaluation, the density value is directly used as the prediction generated from the training set. 10-folds cross validation is applied in the following testings. Larger cross validation results means larger mean accuracy. Table 5.1 and 5.2 compare the cross validation results in the two crime incident datasets. In Table 5.1, it can be observed that the difference is more obvious in cases with short bandwidths because of smaller smoothing ranges. In Table 5.2, the difference is much smaller. This can be explained by the survey by Hipp and Kim which states that crime distribution patterns differ greatly across cities in varying sizes [127]. Another interesting observation is that NKDE gets much better performance on noise incidents under small bandwidths despite the fact that noise disperses along the straight line. One explanation can be that many streets and buildings have screening effect for noises, which is ignored in spatial KDE.

Another metric is hit rate, which is the percentage of incidents that fall within the hotspot areas. However, this metric varies linearly according to the bandwidth so it is not appropriate to be used in evaluation. The second metric is Prediction Accuracy Index (PAI). PAI is defined as:

$$PAI = \frac{\frac{n'}{N}}{\frac{A'}{A}} \tag{5.14}$$

where n' is the number of incidents in hotspot areas, N is the total number of incidents, and A' is the area of hotspots and A is the total area of investigated area.

	Robbery	Noise	Burglary	Public Intoxication
0.2 miles	0.2457	0.5707	0.2185	0.207
30 seconds	0.16069	0.419987	0.141158	0.134438
0.35 miles	0.2335	0.5715	0.2201	0.2072
60 seconds	0.2475	0.5704	0.219	0.2073

Table 5.1: The Cross Validation Results of Crime Incidents Under Four Categories in Tippecanoe County, IN.

	Gangster Disciples	Black P Stones	Black Disciples	Latin Kings
0.4 miles	0.347527	0.24725	0.292861	0.395693
60 seconds	0.347431	0.2471	0.292859	0.39607

Table 5.2: The Cross Validation Results of Gang Member Arrest Records in Chicago, IL.

Table 5.3 and 5.4 list the comparisons of PAI values of two KDEs with different datasets and bandwidth settings. The two tables show higher PAI accuracy from projected spatial KDE than NKDE. However, PAI is highly effected by the choice of bandwidth. In most cases, larger bandwidth leads to smaller PAI [134]. Because of the nonuniform bandwidth of NKDE, it can cover a larger area in some urban areas, which decreases the PAI accuracy. To alleviate this issue, temporal similarity theory is applied to develop another metric.

Another metric is Recapture Rate Index (RRI). This metric is used to evaluate the temporal robustness of hotspot identification. Based on the spatiotemporal near repeats theory in predictive policing, this value should be close to 1 in the ideal scenario. RRI can be defined as:

$$RRI_t = \frac{PAI_t}{PAI_{t-1}} \tag{5.15}$$

	Robbery	Noise	Burglary	Public Intoxication
0.2 miles	88.146	141.79	37.038	155.168
30 seconds	87.278	66.153	43.4928	125.489
0.35 miles	27.277	71.511	15.466	63.071
60 seconds	41.375	36.236	13.083	41.35

Table 5.3: PAI Values of Projected KDE and NKDE for Four Categories in Tippecanoe County, IN.

	Gangster Disciples	Black P Stones	Black Disciples	Latin Kings
0.4 miles	17.57	152.989	114.998	72.19
60 seconds	15.915	82.513	52.551	53.635

Table 5.4: PAI Values of Projected KDE and NKDE for Chicago Gang Member Arrest Records.

where RRI compares *PAI* in the current time frame and the previous time frame. RRI indexes for two datasets are compared in Figure 5.17 and 5.18. The red lines denote the projected KDE and the blue lines denote NKDE. The horizontal lines mark the mean values of the corresponding RRI. In Figure 5.17, the one-year data is divided into 4 three-month time frames. In Figure 5.18, the data is divided into 6 six-month time frames. In the first dataset, it can be observed that KDE gets higher RRI values than NKDE. However, NKDE shows more stability. The differences between mean RRI varies across different crime categories. This indicates different seasonal recapture patterns. In the second dataset, little difference can be observed between the two methods because they are all under the same crime types.

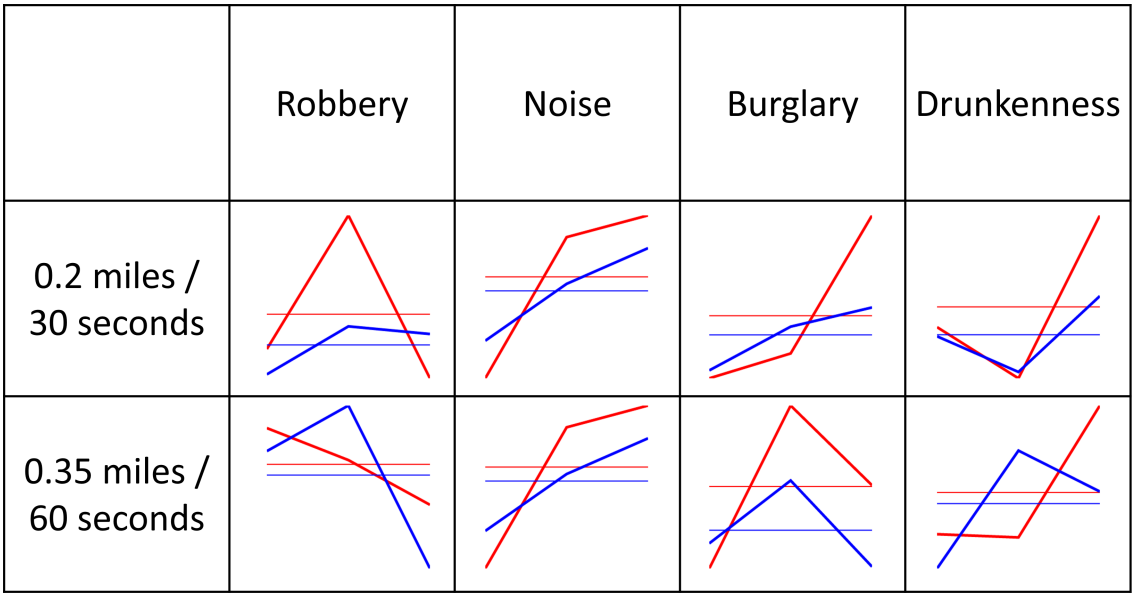


Figure 5.17: RRI Values of Projected KDE and NKDE for Four Categories of Crime Incidents in Tippecanoe County, IN.

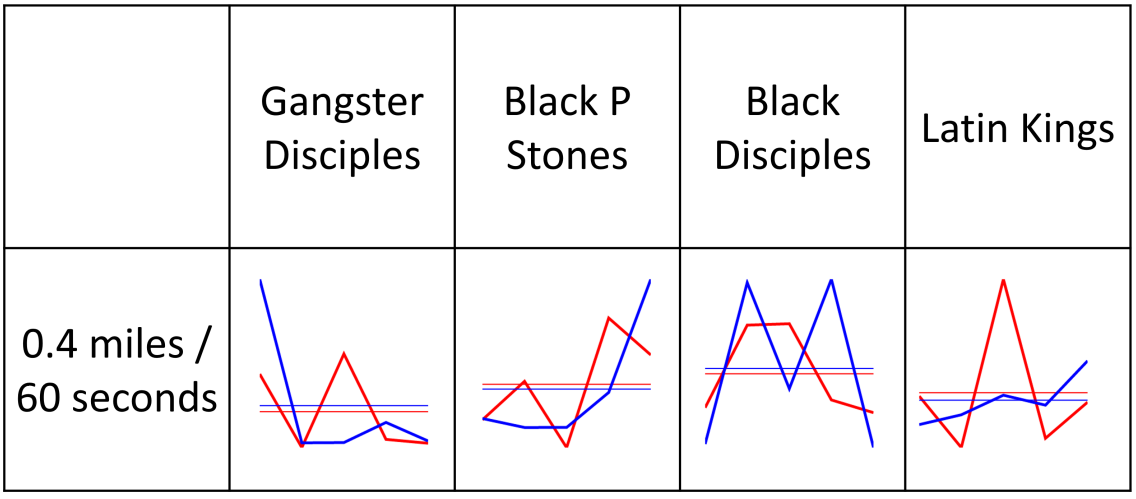


Figure 5.18: RRI Values of Projected KDE and NKDE for Gang Member Arrest Records in Chicago, IL.

Chapter 6

CONCLUSIONS AND FUTURE WORK

To summarize, geographical networked phenomena are analyzed in two visual analytics frameworks. The major contributions of this work include:

- The design of a visual analytics framework for geographical social networks,
- An application of network clustering in geographical network analysis,
- Comparisons of network clustering algorithms,
- An estimation algorithm of event boundaries according to geographical network information, and
- Quantitative metrics for KDE-based hotspot identification methods.

A series of case studies are performed to demonstrate the effectiveness of these frameworks. The frameworks are evaluated with objective evaluations in form of domain expert surveys and quantitative evaluations.

Overall, two visual analytics frameworks are proposed for the analysis of geographical networked data. In the first, virtual networks are projected into their underlying geographical contexts. This framework enables users to discover relationships between the virtual network structures and the geographical properties. The second framework integrates physical network into the spatial hotspot detection algorithm. A boundary extraction algorithm is also proposed to detect the boundaries and overlaps between different categories of events. Besides the physical networks, users can also interactively add guidance to the boundary extraction procedure.

The first part of this thesis proposes a visual analytics framework for virtual to physical network analysis. The interactions in the virtual network space, such as social network, are projected into the underlying geographical space. Through spatial aggregation, interactions between individuals in social networks are projected into interactions between spatial units. The network communities are analyzed through visual analytics. Central nodes are highlighted in the visualizations. The traditional and spatial-near network clustering algorithms are compared in various metrics and visualizations. With hierarchical structures of network clustering algorithms, the relationships between subcomponents in the clusters are also investigated with corresponding visual analytics interfaces. The retweet network in the topic of United States entrepreneurship is used as the case study.

This thesis also proposes a novel framework for visualizing hotspots using the underlying geographic network data. The differences between visualizing data with a traditional two dimensional kernel density estimation and a network based density estimation have been demonstrated. In the cases where geographic network data is available and meaningful to the data set under analysis, the application of network data to geographic hotspot visualization can provide improved contextual cues for analysts. The major design principle that can be taken from this paper is that simple Euclidean based and areal aggregation methods often hide structure in the data. As such, depending on the data properties, aggregation methods that can better encode geographical structure should be employed. Decisions should be made with respect to the analytical task, as well as the need for privacy preservation, as adding more detail can also quickly reveal unique identities in sparse data.

A novel territory extraction method that utilizes level sets of the NKDE along with the DBSCAN algorithm to identify territory boundaries with respect to the underlying geographic network has been introduced. This technique was further augmented by enabling users to interactively add buffer zones for network cuts in the map. Future work will explore the sensitivity of the extracted territories to edge cuts and develop methods to suggest

geographic features that may serve as buffer zones. It is also planned to incorporate point-of-interest datasets to enable more spatial correlation. Currently, NKDE uses the street speed limit as the bandwidth scaling parameter. This assumption can be too strong for certain type of events. it is also planed to develop automatic methods to suggest appropriate edge width settings.

Limitations in the current systems include data scalability, temporal dynamics, and interactivity. In network data exploration, the network size restricts the effectiveness of visualizations. To solve this issue, interactive filtering, multiple linked views, and more advanced graph mining algorithms can be introduced. Temporal dynamics is also interesting in network analysis. This thesis discusses the temporal dynamics in spectral embedding. A future direction can be exploring the relationships between modularity optimization and temporal dynamics analysis. Currently, the network analytics system only has very limited interactions. A future direction is to add more interaction features, such as filtering and temporal dynamics exploration. In hotspot exploring, scalability issues arise when the search space grows. Currently, the algorithm depends on Dijkstra's algorithm which has the complexity $O(|E|+|V|\log|V|)$. Future work could include more advanced navigation algorithms and detail-on-demand estimation schemes. Currently, the system only analyzes the static events within a given time frame. The next feature is to add temporal trend analysis. More interactions can also be introduced to increase spatio-temporal patterns. One example is to extract different types of hotspots to provide more meaningful guidance to the users.

Another major challenge in this work is the appropriate visualization of sets on maps. While set visualization has been a continual area of research [236], methods for linking geographic sets are still in further need of refinement. Future work will explore more advanced methods for set analysis and visualization of categorical variables. The addition of point-of-interest datasets will also benefit from these techniques as relationships between

the proximity of crimes to geographic features (such as convenience stores, bus stations, etc.) are often used when developing spatial models. Besides hotspot maps and territory maps, the spatial distribution can also show different types of attractors, such as point attractors and street attractors [113]. A future feature is to combine different hotspot visualizations together with the temporal information for better support for decision makers.

REFERENCES

[1] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic Geography*, vol. 46, pp. 234–240, 1970.

[2] F. Wang, E. A. Mack, and R. Maciewjewski, "Analyzing entrepreneurial social networks with big data," *Annals of the American Association of Geographers*, vol. 107, no. 1, pp. 130–150, Jan. 2017.

[3] J. J. Thomas and K. A. Cook, Eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Los Alamitos, Calif.: National Visualization and Analytics Ctr, 2005.

[4] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual analytics: Definition, process, and challenges," in *Information Visualization*. Springer, Berlin, Heidelberg, 2008, pp. 154–175.

[5] M. F. Goodchild, "The quality of big (geo)data," *Dialogues in Human Geography*, vol. 3, no. 3, pp. 280–284, Nov. 2013.

[6] ——, "Citizens as sensors: The world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, Nov. 2007.

[7] A. Turner, *Introduction to Neogeography*, 1st ed. O'Reilly, 2006.

[8] A. K. Ziliaskopoulos and S. T. Waller, "An internet-based geographic information system that integrates data, models and users for transportation applications," *Transportation Research Part C: Emerging Technologies*, vol. 8, no. 16, pp. 427–444, Feb. 2000.

[9] T. R. Baker, "Internet-based gis mapping in support of K-12 education," *The Professional Geographer*, vol. 57, no. 1, pp. 44–50, Feb. 2005.

[10] M. Haklay, A. Singleton, and C. Parker, "Web mapping 2.0: The neogeography of the GeoWeb," *Geography Compass*, vol. 2, no. 6, pp. 2011–2039, Nov. 2008.

[11] M. Graham, S. A. Hale, M. Stephens, and V. Mayer-Schönberger, "Geographies of the worlds knowledge," Tech. Rep., 2011. [Online]. Available: http://bit.ly/24Fr9Zm

[12] L. Hollenstein and R. Purves, "Exploring place through user-generated content: Using Flickr tags to describe city cores," *Journal of Spatial Information Science*, vol. 2010, no. 1, pp. 21–48, Jul. 2010.

[13] M. Wall and T. Kirdnark, "Online maps and minorities: Geotagging Thailand's Muslims," *New Media & Society*, vol. 14, no. 4, pp. 701–716, Jun. 2012.

[14] M. Graham and M. Zook, "Visualizing global cyberscapes: Mapping user-generated placemarks," *Journal of Urban Technology*, vol. 18, no. 1, pp. 115–132, Jan. 2011.

[15] M. Crutcher and M. Zook, "Placemarks and waterlines: Racialized cyberscapes in post-Katrina Google Earth," *Geoforum*, vol. 40, no. 4, pp. 523–534, Jul. 2009.

[16] M. Graham, S. A. Hale, and D. Gaffney, "Where in the world are you? Geolocation and language identification in Twitter," *The Professional Geographer*, vol. 66, no. 4, pp. 568–578, Oct. 2014.

[17] M.-P. Kwan, "Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge," *Annals of the American Association of Geographers*, vol. 106, no. 2, pp. 274–282, Mar. 2016.

[18] D. Arribas-Bel, "Accidental, open and everywhere: Emerging data sources for the understanding of cities," *Applied Geography*, vol. 49, pp. 45–53, May 2014.

[19] D. Arribas-Bel and E. Tranos, "New approaches to measure the spatial structure(s) of cities," in *In Proceedings of 23rd GIS Research UK (GISRU) conference*, 2015. [Online]. Available: http://bit.ly/2a5YzJn

[20] M. Batty, "Big data, smart cities and city planning," *Dialogues in Human Geography*, vol. 3, no. 3, pp. 274–279, Nov. 2013.

[21] ——, "Smart cities, big data," *Environment and Planning B: Planning and Design*, vol. 39, no. 2, pp. 191–193, Apr. 2012.

[22] T. Louail, M. Lenormand, O. G. Cantu Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, "From mobile phone data to the spatial structure of cities," *Scientific Reports*, vol. 4, Jun. 2014.

[23] H. Bathelt and J. Glückler, "Toward a relational economic geography," *Journal of Economic Geography*, vol. 3, no. 2, pp. 117–144, Apr. 2003.

[24] E. Tranos and A. Gillespie, "The urban geography of internet backbone networks in europe: Roles and relations," *Journal of Urban Technology*, vol. 18, no. 1, pp. 35–50, Jan. 2011.

[25] H. W.-c. Yeung, "Rethinking relational economic geography," *Transactions of the Institute of British Geographers*, vol. 30, no. 1, pp. 37–51, Mar. 2005.

[26] P. J. Taylor, "Hierarchical tendencies amongst world cities: A global research proposal," *Cities*, vol. 14, no. 6, pp. 323–332, Dec. 1997.

[27] S. Sassen, *The Global City: New York, London, Tokyo*, 2nd ed. Princeton, N.J: Princeton University Press, Sep. 2001.

[28] M. A. Zook and S. D. Brunn, "From podes to antipodes: Positionalities and global airline geographies," *Annals of the Association of American Geographers*, vol. 96, no. 3, pp. 471–490, Sep. 2006.

[29] E. Tranos and P. Nijkamp, "The death of distance revisited: Cyber-Place, physical and relational proximities," *Journal of Regional Science*, vol. 53, no. 5, pp. 855–873, Dec. 2013.

[30] H. E. Aldrich and C. Zimmer, "Entrepreneurship through social networks," in *California Management Review*, 1986, vol. 33, pp. 3–23.

[31] S. Birley, "The role of networks in the entrepreneurial process," *Journal of Business Venturing*, vol. 1, no. 1, pp. 107–117, Dec. 1985.

[32] D. M. Sullivan and C. M. Ford, "How entrepreneurs use networks to address changing resource requirements during early venture development," *Entrepreneurship Theory and Practice*, vol. 38, no. 3, pp. 551–574, May 2014.

[33] J. Bruderl and P. Preisendorfer, "Network support and the success of newly founded business," *Small Business Economics*, vol. 10, no. 3, pp. 213–225, 1998.

[34] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[35] T. Elfring and W. Hulsink, "Networking by entrepreneurs: Patterns of tie-formation in emerging organizations," *Organization Studies*, vol. 28, no. 12, pp. 1849–1872, Dec. 2007.

[36] S. L. Jack, "The role, use and activation of strong and weak network ties: A qualitative analysis," *Journal of Management Studies*, vol. 42, no. 6, pp. 1233–1259, Sep. 2005.

[37] P. Dubini and H. E. Aldrich, "Personal and extended networks are central to the entrepreneurial process," *Journal of Business Venturing*, vol. 6, pp. 305–313, 1991.

[38] E. Fischer and A. R. Reuber, "Social interaction via new social media: (How) can interactions on Twitter affect effectual thinking and behavior?" *Journal of Business Venturing*, vol. 26, no. 1, pp. 1–18, Jan. 2011.

[39] J. van der Krogt, "The influence of social media on nascent entrepreneurship in the Netherlands," Master Thesis, Tilburg University, 2011.

[40] J. Stevens, "Shattering the boundaries through self-efficacy : Exploring the social media habits of South African previously disadvantaged entrepreneurs," MPhill Thesis, Stellenbosch University, Mar. 2013.

[41] A. Saxenian, *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*, 50525th ed. Cambridge, Mass.: Harvard University Press, Mar. 1996.

[42] ——, *The New Argonauts: Regional Advantage in a Global Economy*, first edition ed. Cambridge, Mass.: Harvard University Press, Oct. 2007.

[43] M. Granovetter, "The strength of weak ties: A network theory revisited," *Sociological Theory*, vol. 1, pp. 201–233, 1983.

[44] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu, "Whisper: Tracing the spatiotemporal process of information diffusion in real time," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2649 –2658, Dec. 2012.

[45] I. Taxidou and P. M. Fischer, "RApID: A system for real-time analysis of information diffusion in Twitter," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2014, pp. 2060–2062.

[46] C. Wang, Z. Xiao, Y. Liu, Y. Xu, A. Zhou, and K. Zhang, "SentiView: Sentiment analysis and visualization for internet popular topics," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 6, pp. 620–630, Nov. 2013.

[47] E. M. Cody, A. J. Reagan, L. Mitchell, P. S. Dodds, and C. M. Danforth, "Climate change sentiment on Twitter: An unsolicited public opinion poll," *PLOS ONE*, vol. 10, no. 8, p. e0136092, Aug. 2015.

[48] N. Cao, L. Lu, Y.-R. Lin, F. Wang, and Z. Wen, "SocialHelix: Visual analysis of sentiment divergence in social media," *Journal of Visualization*, vol. 18, no. 2, pp. 221–235, May 2015.

[49] J. Zhao, N. Cao, Z. Wen, Y. Song, Y. R. Lin, and C. Collins, "#FluxFlow: Visual analysis of anomalous information spreading on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1773–1782, Dec. 2014.

[50] P. Xu, Y. Wu, E. Wei, T. Q. Peng, S. Liu, J. J. H. Zhu, and H. Qu, "Visual analysis of topic competition on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2012–2021, Dec. 2013.

[51] G. Sun, Y. Wu, S. Liu, T. Q. Peng, J. J. H. Zhu, and R. Liang, "EvoRiver: Visual analysis of topic coopetition on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1753–1762, Dec. 2014.

[52] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu, "OpinionFlow: Visual analysis of opinion diffusion on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1763–1772, Dec. 2014.

[53] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York, NY, USA: ACM, 2013, pp. 344–353.

[54] S. Chen, X. Yuan, Z. Wang, C. Guo, J. Liang, Z. Wang, X. L. Zhang, and J. Zhang, "Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 270–279, Jan. 2016.

[55] H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl, "ScatterBlogs2: Real-time monitoring of microblog messages through user-guided filtering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2022–2031, Dec. 2013.

[56] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "SensePlace2: GeoTwitter analytics support for situational awareness," in *IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct. 2011, pp. 181–190.

[57] J. Chae, D. Thom, H. Bosch, J. Yang, R. Maciejewski, D. S. Ebert, and T. Ertl, "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition," in *IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct. 2012.

[58] J. Chae, Y. Cui, Y. Jang, G. Wang, A. Malik, and D. S. Ebert, "Trajectory-based visual analytics for anomalous human movement analysis using social media," in *EuroVis Workshop on Visual Analytics (EuroVA)*, E. Bertini and J. C. Roberts, Eds. The Eurographics Association, 2015.

[59] D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl, "Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages," in *IEEE Pacific Visualization Symposium*, Feb. 2012, pp. 41–48.

[60] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, "LeadLine: Interactive visual analysis of text data through event identification and exploration," in *IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct. 2012, pp. 93–102.

[61] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski, "#Earthquake: Twitter as a distributed sensor system," *Transactions in GIS*, vol. 17, no. 1, pp. 124–147, Feb. 2013.

[62] K. C. Cox, S. G. Eick, and T. He, "3d geographic network displays," *SIGMOD Rec.*, vol. 25, no. 4, pp. 50–54, Dec. 1996.

[63] B. Alper, S. Smengen, and S. Balcisoy, "Dynamic visualization of geographic networks using surface deformations with constraints," in *Proceedings of the Computer Graphics International Conference (CGI), Computer Graphics Society, Petrópolis, Brazil*, 2007.

[64] W. Cui, H. Zhou, H. Qu, P. C. Wong, and X. Li, "Geometry-based edge clustering for graph visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1277–1284, Nov. 2008.

[65] A. Lambert, R. Bourqui, and D. Auber, "3d edge bundling for geographical data visualization," in *14th International Conference Information Visualisation*, Jul. 2010, pp. 329–335.

[66] D. Holten, "Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 741–748, Sep. 2006.

[67] D. Holten and J. J. van Wijk, "Force-directed edge bundling for graph visualization," in *Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization*. Chichester, UK: The Eurographs Association & John Wiley & Sons, Ltd., 2009, pp. 983–998.

[68] C. Hurter, O. Ersoy, and A. Telea, "Graph bundling by kernel density estimation," *Computer Graphics Forum*, vol. 31, no. 3pt1, pp. 865–874, Jun. 2012.

[69] P. Rodgers, "Graph drawing techniques for geographic visualization," in *Exploring geovisualization*, A. M. MacEachren, M.-J. Kraak, and J. Dykes, Eds. Pergamon, Dec. 2004, pp. 143–158.

[70] R. Tamassia, Ed., *Handbook of Graph Drawing and Visualization*, 1st ed. Boca Raton: Chapman and Hall/CRC, Aug. 2013.

[71] A. M. Voorhees, "A general theory of traffic movement," *Transportation*, vol. 40, no. 6, pp. 1105–1116, Nov. 2013.

[72] J. Wood, J. Dykes, and A. Slingsby, "Visualisation of origins, destinations and flows with OD maps," *The Cartographic Journal*, vol. 47, no. 2, pp. 117–129, May 2010.

[73] Y. Yang, T. Dwyer, S. Goodwin, and K. Marriott, "Many-to-many geographically-embedded flow visualisation: An evaluation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 411–420, Jan. 2017.

[74] G. Andrienko and N. Andrienko, "Spatio-temporal aggregation for visual analysis of movements," in *IEEE Symposium on Visual Analytics Science and Technology*, Oct. 2008, pp. 51–58.

[75] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko, "Stacking-based visualization of trajectory attribute data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2565–2574, Dec. 2012.

[76] R. Scheepens, N. Willems, H. van de Wetering, and J. van Wijk, "Interactive density maps for moving objects," *IEEE Computer Graphics and Applications*, vol. 32, no. 1, pp. 56–66, 2012.

[77] T. Cheng, G. Tanaksaranond, C. Brunsdon, and J. Haworth, "Exploratory visualisation of congestion evolutions on urban transport networks," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 296–306, Nov. 2013.

[78] G. Sun, Y. Liu, W. Wu, R. Liang, and H. Qu, "Embedding temporal display into maps for occlusion-free visualization of spatio-temporal data," in *IEEE Pacific Visualization Symposium*, Mar. 2014, pp. 185–192.

[79] G. Sun, R. Liang, H. Qu, and Y. Wu, "Embedding spatio-temporal information into maps by route-zooming," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, no. 99, pp. 1–1, 2016.

[80] S. Kim, R. Maciejewski, A. Malik, Y. Jang, D. Ebert, and T. Isenberg, "Bristle maps: A multivariate abstraction technique for geovisualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 9, pp. 1438–1454, 2013.

[81] N. Wong, S. Carpendale, and S. Greenberg, "Edgelens: An interactive method for managing edge congestion in graphs," in *IEEE Symposium on Information Visualization*, Oct. 2003, pp. 51–58.

[82] R. Krüger, D. Thom, M. Wörner, H. Bosch, and T. Ertl, "TrajectoryLenses A set-based filtering and exploration technique for long-term trajectory data," *Computer Graphics Forum*, vol. 32, pp. 451–460, Jun. 2013.

[83] H. Liu, Y. Gao, L. Lu, S. Liu, H. Qu, and L. M. Ni, "Visual analysis of route diversity," in *IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct. 2011, pp. 171–180.

[84] H. Hu, H. Zhang, and W. Li, "Visualizing network communication in geographic environment," in *International Conference on Virtual Reality and Visualization*, Sep. 2013, pp. 206–212.

[85] Z. Wang, T. Ye, M. Lu, X. Yuan, H. Qu, J. Yuan, and Q. Wu, "Visual exploration of sparse traffic trajectory data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1813–1822, Dec. 2014.

[86] J. Leskovec, K. J. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection," in *Proceedings of the 19th international conference on World wide web*. New York, NY, USA: ACM, 2010, pp. 631–640.

[87] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.

[88] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, p. 026113, Feb. 2004.

[89] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, Mar. 2004.

[90] T. Leighton and S. Rao, "An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms," in *Proceedings of 29th Annual Symposium on Foundations of Computer Science*, Oct. 1988, pp. 422–431.

[91] R. Andersen and K. J. Lang, "Communities from seed sets," in *Proceedings of the 15th International Conference on World Wide Web*. New York, NY, USA: ACM, 2006, pp. 223–232.

[92] S. E. Schaeffer, "Survey: Graph clustering," *Comput. Sci. Rev.*, vol. 1, no. 1, pp. 27–64, Aug. 2007.

[93] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad and spectral," *J. ACM*, vol. 51, no. 3, pp. 497–515, May 2004.

[94] G. W. Flake, R. E. Tarjan, and K. Tsioutsiouliklis, "Graph clustering and minimum cut trees," *Internet Mathematics*, vol. 1, no. 4, pp. 385–408, 2003.

[95] G. W. Flake, S. Lawrence, and C. L. Giles, "Efficient identification of web communities," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2000, pp. 150–160.

[96] D. L. Wallace, "Comment," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 569–576, Sep. 1983.

[97] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, p. 066133, Jun. 2004.

[98] ——, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006.

[99] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in *Proceedings of the 2005 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Apr. 2005, pp. 274–285.

[100] X. Xiong, J. Ma, M. Wang, G. Zhou, and K. Xu, "Information diffusion model in modular microblogging networks," *World Wide Web*, vol. 18, no. 4, pp. 1051–1069, Jul. 2015.

[101] T. C. Bailey and A. C. Gatrell, *Interactive spatial data analysis*. Harlow Essex, England; New York, NY: Longman Scientific & Technical ; J. Wiley, 1995.

[102] J. Eck, S. Chainey, J. Cameron, and R. Wilson, "Mapping crime: Understanding hotspots," National Institute of Justice, Washington DC, Report, Aug. 2005.

[103] D. O'Sullivan and D. Unwin, *Geographic Information Analysis*, 2nd ed. Hoboken, N.J: Wiley, Mar. 2010.

[104] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, Apr. 1986.

[105] N. Koutsias, P. Balatsos, and K. Kalabokidis, "Fire occurrence zones: Kernel density estimation of historical wildfire ignitions at the national level, Greece," *Journal of Maps*, vol. 10, no. 4, pp. 630–639, 2014.

[106] O. D. Lampe and H. Hauser, "Interactive visualization of streaming data with kernel density estimation," in *IEEE Pacific Visualization Symposium (PacificVis)*, Mar. 2011, pp. 171–178.

[107] N. Levine, "CrimeStat: A spatial statistical program for the analysis of crime incidents," in *Encyclopedia of GIS*. Springer US, 2008, pp. 187–193.

[108] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert, "A visual analytics approach to understanding spatiotemporal hotspots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 2, pp. 205–220, Mar. 2010.

[109] R. Scheepens, H. van de Wetering, and J. J. van Wijk, "Contour based visualization of vessel movement predictions," *International Journal of Geographical Information Science*, vol. 28, no. 5, pp. 891–909, 2014.

[110] R. Tao and J.-C. Thill, "Spatial cluster detection in spatial flow data," *Geographical Analysis*, vol. 48, no. 4, pp. 355–372, Nov. 2016.

[111] A. Razip, A. Malik, S. Afzal, M. Potrawski, R. Maciejewski, Y. Jang, N. Elmqvist, and D. Ebert, "A mobile visual analytics approach for law enforcement situation awareness," in *IEEE Pacific Visualization Symposium (PacificVis)*, 2014, pp. 169–176.

[112] J. M. Krisp and S. Peters, "Directed kernel density estimation (DKDE) for time series visualization," *Annals of GIS*, vol. 17, no. 3, pp. 155–162, 2011.

[113] J. H. Ratcliffe, "The hotspotmatrix: A framework for the spatiotemporal targeting of crime reduction," *Police Practice and Research*, vol. 5, no. 1, pp. 5–23, Mar. 2004.

[114] H. J. Miller, "Potential contributions of spatial analysis to geographic information systems for transportation (GIS-T)," *Geographical Analysis*, vol. 31, no. 4, pp. 373–399, 1999.

[115] Z. Xie and J. Yan, "Kernel density estimation of traffic accidents in a network space," *Computers, Environment and Urban Systems*, vol. 32, no. 5, pp. 396 – 406, 2008.

[116] A. Okabe, *Spatial Analysis Along Networks: Statistical and Computational Methods*, 1st ed. Wiley, 2012.

[117] A. Okabe, K.-i. Okunuki, and S. Shiode, "SANET: A toolbox for spatial analysis on a network," *Geographical Analysis*, vol. 38, no. 1, pp. 57–66, 2006.

[118] A. Okabe and I. Yamada, "The K-function method on a network and its computational implementation," *Geographical Analysis*, vol. 33, no. 3, pp. 271–290, 2001.

[119] G. Borruso, "Network density estimation: A GIS approach for analysing point patterns in a network space," *Transactions in GIS*, vol. 12, no. 3, pp. 377–402, 2008.

[120] S. Shiode and N. Shiode, "Detection of multi-scale clusters in network space," *International Journal of Geographical Information Science*, vol. 23, no. 1, pp. 75–92, 2009.

[121] L. Tompson, H. Partridge, and N. Shepherd, "Hot routes: Developing a new technique for the spatial analysis of crime," *Crime Mapping: A Journal of Research and Practice*, vol. 1, no. 1, pp. 77–96, 2009.

[122] K. Heim, "Visualization and modeling for crime data indexed by road segments," Ph.D. dissertation, George Mason University, 2014.

[123] S. W. Laffan and M. D. Taylor, "FishTracker: A GIS toolbox for kernel density estimation of animal home ranges that accounts for transit times and hard boundaries," in *Proceedings of 20th International Congress on Modelling and Simulation*, 2013.

[124] N. Ferreira, L. Lins, D. Fink, S. Kelling, C. Wood, J. Freire, and C. Silva, "BirdVis: Visualizing and understanding bird populations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2374–2383, Dec. 2011.

[125] J. E. Eck and D. Weisburd, Eds., *Crime and Place: Crime Prevention Studies*. Monsey, N.Y.; Washington, D.C.: Willow Tree Pr, Dec. 1995.

[126] M. A. Andresen, S. J. Linning, and N. Malleson, "Crime at places and spatial concentrations: Exploring the spatial stability of property crime in Vancouver, BC, 20032013," *Journal of Quantitative Criminology*, pp. 1–21, Mar. 2016.

[127] J. R. Hipp and Y.-A. Kim, "Measuring crime concentration across cities of varying sizes: Complications based on the spatial and temporal scale employed," *Journal of Quantitative Criminology*, pp. 1–38, Nov. 2016.

[128] L. W. Sherman and D. Weisburd, "General deterrent effects of police patrol in crime "hot spots": A randomized, controlled trial," *Justice Quarterly*, vol. 12, no. 4, pp. 625–648, Dec. 1995.

[129] J. H. Ratcliffe, T. Taniguchi, E. R. Groff, and J. D. Wood, "The Philadelphia foot patrol experiment: A randomized controlled trial of police patrol effectiveness in violent crime hotspots," *Criminology*, vol. 49, no. 3, pp. 795–831, Aug. 2011.

[130] A. A. Braga, A. V. Papachristos, and D. M. Hureau, "The effects of hot spots policing on crime: An updated systematic review and meta-analysis," *Justice Quarterly*, vol. 31, no. 4, pp. 633–663, Jul. 2014.

[131] B. Ariel and H. Partridge, "Predictable policing: Measuring the crime control benefits of hotspots policing at bus stops," *Journal of Quantitative Criminology*, pp. 1–25, Jun. 2016.

[132] B. Taylor, C. S. Koper, and D. J. Woods, "A randomized controlled trial of different policing strategies at hot spots of violent crime," *Journal of Experimental Criminology*, vol. 7, no. 2, pp. 149–181, Jun. 2011.

[133] E. R. Groff, J. H. Ratcliffe, C. P. Haberman, E. T. Sorg, N. M. Joyce, and R. B. Taylor, "Does what police do at hot spots matter? The Philadelphia policing tactics experiment," *Criminology*, vol. 53, no. 1, pp. 23–53, Feb. 2015.

[134] T. C. Hart and P. A. Zandbergen, "Effects of data quality on predictive hotspot mapping," U.S. Department of Justice, Tech. Rep. 239861, Oct. 2012.

[135] S. Chainey, L. Tompson, and S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime," *Security Journal*, vol. 21, no. 1-2, pp. 4–28, Feb. 2008.

[136] N. Levine, "The 'hottest' part of a hotspot: Comments on 'the utility of hotspot mapping for predicting spatial patterns of crime'," *Secuity Journal*, vol. 21, no. 4, pp. 295–302, Oct. 2008.

[137] G. Drawve, S. C. Moak, and E. R. Berthelot, "Predictability of gun crimes: A comparison of hot spot and risk terrain modelling techniques," *Policing and Society*, vol. 26, no. 3, pp. 312–331, Apr. 2016.

[138] F. Morstatter, H. Gao, and H. Liu, "Discovering location information in social media," *IEEE Computer Science, Data Engineering Bulletin*, vol. 38, no. 2, pp. 4–13, 2015.

[139] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, Apr. 2015.

[140] R. Kitchin, "Big data and human geography opportunities, challenges and risks," *Dialogues in Human Geography*, vol. 3, no. 3, pp. 262–267, Nov. 2013.

[141] D. Boyd and K. Crawford, "Critical questions for big data," *Information, Communication & Society*, vol. 15, no. 5, pp. 662–679, Jun. 2012.

[142] A. Mislove, S. L. Jørgensen, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of Twitter users," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. AAAI Press, 2011, pp. 554–557.

[143] J. W. Crampton, M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. W. Wilson, and M. Zook, "Beyond the geotag: Situating 'big data' and leveraging the potential of the geoweb," Social Science Research Network, Rochester, NY, Tech. Rep., Apr. 2013.

[144] A. Greve and J. W. Salaff, "Social networks and entrepreneurship," *Entrepreneurship Theory and Practice*, vol. 28, no. 1, pp. 1–22, Sep. 2003.

[145] B. Feld and D. Kaplan, *Startup Communities: Building an Entrepreneurial Ecosystem in Your City*, unabridged edition ed. Brilliance Audio, Sep. 2013.

[146] K. Brady, "20 Twitter hashtags that will turn you into an entrepreneurial rock star," 2016. [Online]. Available: http://bit.ly/1s2qUEA

[147] StartupSmart, "Infographic: Twitter hashtags for entrepreneurs," Oct. 2013. [Online]. Available: http://www.startupsmart.com.au/advice/sales-and-marketing/infographic-twittter-hashtags-for-entrepreneurs/

[148] S. White, "45 popular entrepreneurial Twitter hashtags you should use today," Apr. 2015. [Online]. Available: https://www.socialquant.net/popular-twitter-hashtags/

[149] M. Cha, F. Benevenuto, H. Haddadi, and K. Gummadi, "The world of connections and information flow in Twitter," *Trans. Sys. Man Cyber. Part A*, vol. 42, no. 4, pp. 991–998, Jul. 2012.

[150] A. Croitoru, A. Crooks, J. Radzikowski, and A. Stefanidis, "Geosocial gauge: A system prototype for knowledge discovery from social media," *International Journal of Geographical Information Science*, vol. 27, no. 12, pp. 2483–2508, Dec. 2013.

[151] P. H. Wilken, *Entrepreneurship: A Comparative and Historical Study*. Ablex Publishing Corporation, Jan. 1979.

[152] Inc., "The 2015 Inc. 5000," Inc.com, Tech. Rep., 2015. [Online]. Available: http://www.inc.com/inc5000

[153] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? Comparing data from Twitters streaming API with Twitters firehose," in *International Conference on Weblogs and Social Media*. AAAI, 2013, pp. 400–408.

[154] Twitter Inc., "Twitter search API," 2011. [Online]. Available: https://dev.twitter.com/rest/public/search

[155] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2011, pp. 237–246.

[156] J. Sampson, F. Morstatter, R. Maciejewski, and H. Liu, "Surpassing the limit: Keyword clustering to improve Twitter sample coverage," in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. New York, NY, USA: ACM, 2015, pp. 237–245.

[157] K. Crawford, "Following you: Disciplines of listening in social media," *Continuum*, vol. 23, no. 4, pp. 525–535, Aug. 2009.

[158] T. Inc., "One hundred million voices," 2011. [Online]. Available: http://bit.ly/2ajdB1n

[159] M. Duggan, N. B. Ellison, C. Lampe, A. Lenhart, and M. Madden, "Demographics of key social networking platforms," Jan. 2015. [Online]. Available: http://pewrsr.ch/1xMwvDG

[160] M. Adnan, P. A. Longley, and S. M. Khan, "Social dynamics of Twitter usage in London, Paris, and New York city," *First Monday*, vol. 19, no. 5, May 2014.

[161] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who Tweets? deriving the demographic characteristics of age, occupation and social class from Twitter user metadata," *PLOS ONE*, vol. 10, no. 3, p. e0115545, Mar. 2015.

[162] T. Kautonen, "Understanding the older entrepreneur: Comparing third age and prime age entrepreneurs in Finland," *International Journal of Business Science and Applied Management*, 2008.

[163] B. Rotefoss and L. Kolvereid, "Aspiring, nascent and fledgling entrepreneurs: An investigation of the business start-up process," *Entrepreneurship & Regional Development*, vol. 17, no. 2, pp. 109–127, Mar. 2005.

[164] J. Curran and R. A. Blackburn, "Older people and the enterprise society: Age and self-employment propensities," *Work, Employment and Society*, vol. 15, no. 04, pp. 889–902, Dec. 2001.

[165] M. Hart, M. Anyadike-Danes, and R. Blackburn, "Spatial differences in entrepreneurship: A comparison of prime age and third age cohorts," Economic Research Institute of Northern Ireland, Tech. Rep., 2004. [Online]. Available: http://bit.ly/2abmH1Z

[166] D. Stangler and E. Marion, "The age of the entrepreneur: Demographics and entrepreneurship," in *i4j Summit*, 2013. [Online]. Available: http://bit.ly/1Kb7FDK

[167] R. W. Fairlie, A. Morelix, E. Reedy, and J. Russell, "The Kauffman index of startup activity: National trends," The Kauffman Foundation, Tech. Rep., 2015. [Online]. Available: http://bit.ly/1QcG0RV

[168] Center for Excellence in Service, "The state of small business report December 2009," Tech. Rep., 2010. [Online]. Available: http://bit.ly/2pDYDfe

[169] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[170] F. Morstatter, S. Kumar, H. Liu, and R. Maciejewski, "Understanding twitter data with tweetxplorer," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1482–1485.

[171] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, p. 066111, Dec. 2004.

[172] M. L. Moss and A. M. Townsend, "The internet backbone and the American metropolis," *The Information Society*, vol. 16, no. 1, pp. 35–47, Mar. 2000.

[173] Charlotte Neighborhood and Business Services, "High growth entrepreneurship strategy," Tech. Rep., 2012. [Online]. Available: http://bit.ly/275HqsL

[174] Charlotte Entrepreneur Organization, "Entrepreneurs organization, Charlotte chapter," Tech. Rep., 2016. [Online]. Available: http://www.eocharlotte.org

[175] Charlotte Regional Fund for Entrepreneurship, "Charlotte regional fund for entrepreneurship," Tech. Rep., 2016. [Online]. Available: http://charlotteentrepreneur.org/

[176] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, "Maximizing modularity is hard," *arXiv:physics/0608255*, Aug. 2006.

[177] K. Wakita and T. Tsurumi, "Finding community structure in mega-scale social networks: [Extended Abstract]," in *Proceedings of the 16th International Conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 1275–1276.

[178] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, Oct. 2008.

[179] I. Herman, G. Melançon, M. M. d. Ruiter, and M. Delest, "Latour: A tree visualisation system," in *Graph Drawing*. Springer, Berlin, Heidelberg, Sep. 1999, pp. 392–399.

[180] M. Schonlau, "Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams," *Computational Statistics*, vol. 19, no. 1, pp. 95–111, Feb. 2004.

[181] V. d. F. Vieira, C. R. Xavier, N. F. F. Ebecken, and A. G. Evsukoff, "Performance evaluation of modularity based community detection algorithms in large scale networks," *Mathematical Problems in Engineering*, vol. 2014, p. e502809, Dec. 2014.

[182] X. Liu, T. Murata, and K. Wakita, "Extending modularity by capturing the similarity attraction feature in the null model," *arXiv:1210.4007 [physics]*, Oct. 2012.

[183] J. Hannigan, G. Hernandez, R. M. Medina, P. Roos, and P. Shakarian, "Mining for spatially-near communities in geo-located social networks," in *Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium*, 2013.

[184] P. Shakarian, P. Roos, D. Callahan, and C. Kirk, "Mining for geographically disperse communities in social networks by leveraging distance modularity," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2013, pp. 1402–1409.

[185] M. Meila and D. Heckerman, "An experimental comparison of model-based clustering methods," *Machine Learning*, vol. 42, no. 1-2, pp. 9–29, Jan. 2001.

[186] M. Meila, "Comparing clusterings by the variation of information," in *Learning Theory and Kernel Machines*. Springer, Berlin, Heidelberg, 2003, pp. 173–187.

[187] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 553–569, 1983.

[188] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[189] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985.

[190] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in *Proceedings of Pacific Symposium on Biocomputing*, vol. 7, 2002, pp. 6–17.

[191] B. H. Good, Y.-A. de Montjoye, and A. Clauset, "Performance of modularity maximization in practical contexts," *Physical Review E*, vol. 81, no. 4, p. 046106, Apr. 2010.

[192] R. F. Cohen, P. Eades, T. Lin, and F. Ruskey, "Three-dimensional graph drawing," in *Graph Drawing*. Springer, Berlin, Heidelberg, Oct. 1994, pp. 1–11.

[193] S. Yan, D. Xu, B. Zhang, H. j. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[194] D. B. Skillicorn, Q. Zheng, and C. Morselli, "Spectral embedding for dynamic social networks," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Aug. 2013, pp. 316–323.

[195] T. C. Matisziw, T. H. Grubesic, and J. Guo, "Robustness elasticity in complex networks," *PLOS ONE*, vol. 7, no. 7, p. e39788, Jul. 2012.

[196] C. Chen, H. Tong, B. A. Prakash, C. E. Tsourakakis, T. Eliassi-Rad, C. Faloutsos, and D. H. Chau, "Node immunization on large graphs: Theory and algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 113–126, Jan. 2016.

[197] M. J. F. Alenazi and J. P. G. Sterbenz, "Evaluation and comparison of several graph robustness metrics to improve network resilience," in *7th International Workshop on Reliable Networks Design and Modeling (RNDM)*, Oct. 2015, pp. 7–13.

[198] F. Malliaros, V. Megalooikonomou, and C. Faloutsos, "Fast robustness estimation in large social graphs: Communities and anomaly detection," in *Proceedings of the SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Apr. 2012, pp. 942–953.

[199] M. Benzi and C. Klymko, "Total communicability as a centrality measure," *Journal of Complex Networks*, vol. 1, no. 2, pp. 124–149, Dec. 2013.

[200] A. Perrot, R. Bourqui, N. Hanusse, F. Lalanne, and D. Auber, "Large interactive visualization of density functions on big data infrastructure," in *5th Symposium on Large Data Analysis and Visualization (LDAV)*, Oct. 2015, pp. 99–106.

[201] V. Spicer, J. Song, P. Brantingham, A. Park, and M. A. Andresen, "Street profile analysis: A new method for mapping crime on major roadways," *Applied Geography*, vol. 69, pp. 65 – 74, 2016.

[202] N.-B. Heidenreich, A. Schindler, and S. Sperlich, "Bandwidth selection for kernel density estimation: A review of fully automatic selectors," *AStA Advances in Statistical Analysis*, vol. 97, no. 4, pp. 403–433, 2013.

[203] J. A. Downs and M. W. Horner, "Probabilistic potential path trees for visualizing and analyzing vehicle tracking data," *Journal of Transport Geography*, vol. 23, pp. 72–80, Jul. 2012.

[204] W. Yu and T. Ai, "The visualization and analysis of urban facility pois using network kernel density estimation constrained by multi-factors," *Boletim de Ciłncias Geodsicas*, vol. 20, pp. 902 – 926, 2014.

[205] S. Huddleston, J. Fox, and D. Brown, "Mapping gang spheres of influence," *Crime Mapping: A Journal of Research and Practice*, vol. 4, no. 2, pp. 39–67, 2012.

[206] K. Nakamura, G. Tita, and D. Krackhardt, "Violence in the "balance": A structural analysis of how rivals, allies, and third-parties shape inter-gang violence," *Heinz College Research*, Apr. 2011.

[207] G. Tita, K. J. Riley, G. Ridgeway, and C. Grammich, "Unruly turf: The role of interagency collaborations in reducing gun violence," 2013. [Online]. Available: http://bit.ly/2o5yOVo

[208] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1996, pp. 226–231.

[209] W. M. Getz, S. Fortmann-Roe, P. C. Cross, A. J. Lyons, S. J. Ryan, and C. C. Wilmers, "LoCoH: Nonparameteric kernel methods for constructing home ranges and utilization distributions," *PLOS ONE*, vol. 2, no. 2, p. e207, Feb. 2007.

[210] W. M. Getz and C. C. Wilmers, "A local nearest-neighbor convex-hull construction of home ranges and utilization distributions," *Ecography*, vol. 27, no. 4, pp. 489–505, 2004.

[211] B. J. Worton, "Kernel methods for estimating the utilization distribution in home-range studies," *Ecology*, vol. 70, no. 1, pp. 164–168, 1989.

[212] A. K. Bhattacharya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[213] S. M. Kang and R. P. Wildes, "The n-distribution Bhattacharyya coefficient," York University, Tech. Rep. EECS-2015-02, Feb. 2015.

[214] W. Bernasco and F. Luykx, "Effects of attractiveness, opportunity and accessibility to burglars on residential burglary rates of urban neighborhoods," *Criminology*, vol. 41, no. 3, pp. 981–1002, 2003.

[215] I. Colquhoun, *Design Out Crime*. Routledge, Mar. 2007.

[216] K. D. Rossmo, "Geographic profiling: Target patterns of serial murderers," Ph.D. dissertation, Simon Fraser University, 1987.

[217] A. Malik, R. Maciejewski, E. Hodgess, and D. S. Ebert, "Describing temporal correlation spatially in a visual analytics environment," in *44th Hawaii International Conference on System Sciences (HICSS)*, Jan. 2011, pp. 1–8.

[218] Huffington Post, "Lil JoJo dead: Teen chicago rapper joseph coleman fatally shot, police investigate Chief Keef's Tweets," *Huffington Post*, Sep. 2012. [Online]. Available: http://huff.to/2oooJPH

[219] ——, "William Balfour confessed to killing Jennifer Hudson's family, suspect's mistress claims," *Huffington Post*, Apr. 2012. [Online]. Available: http://huff.to/2oQWIRX

[220] T. Dunong, "Bandwidth selectors for multivariate kernel density estimation," *Bulletin of the Australian Mathematical Society*, vol. 71, no. 02, pp. 351–352, 2005.

[221] Q. Li and J. S. Racine, *Nonparametric econometrics: Theory and practice*. Princeton University Press, 2007.

[222] D. W. Scott, *Multivariate density estimation: Theory, practice, and visualization*. John Wiley & Sons, 2015.

[223] S. J. Sheather, "Density estimation," *Statistical Science*, vol. 19, no. 4, pp. 588–597, Nov. 2004.

[224] M. P. Wand and M. C. Jones, *Kernel smoothing*. Crc Press, 1994.

[225] R. P. W. Duin, "On the choice of smoothing parameters for Parzen estimators of probability density functions," *IEEE Transactions on Computers*, vol. C-25, no. 11, pp. 1175–1179, Nov. 1976.

[226] J. Habbema, J. Hermans, and K. Van den Broek, "A stepwise discrimination analysis program using density estimation," in *Proceedings in Computational Statistics*. Vienna: Physica Verlag, 1974.

[227] A. W. Bowman, P. Hall, and D. M. Titterington, "Cross-validation in nonparametric estimation of probabilities and probability densities," *Biometrika*, vol. 71, no. 2, pp. 341–351, Aug. 1984.

[228] J. S. Horne and E. O. Garton, "Likelihood cross-validation versus least squares cross validation for choosing the smoothing parameter in kernel home-range analysis," *Journal of Wildlife Management*, 2006.

[229] M. Rudemo, "Empirical choice of histograms and kernel density estimators," *Scandinavian Journal of Statistics*, vol. 9, no. 2, pp. 65–78, 1982.

[230] D. W. Scott and G. R. Terrell, "Biased and unbiased cross-validation in density estimation," *Journal of the American Statistical Association*, vol. 82, no. 400, pp. 1131–1146, Dec. 1987.

[231] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 683–690, 1991.

[232] P. Hall, S. J. Sheather, M. C. Jones, and J. S. Marron, "On optimal data-based bandwidth selection in kernel density estimation," *Biometrika*, vol. 78, no. 2, pp. 263–269, Jun. 1991.

[233] S.-T. Chiu, "Bandwidth selection for kernel density estimation," *The Annals of Statistics*, vol. 19, no. 4, pp. 1883–1905, Dec. 1991.

[234] M. Rosenblatt, "Curve estimates," *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1815–1842, 1971.

[235] Y. Zheng and J. M. Phillips, "L$nfty$ error and bandwidth selection for kernel density estimates of large data," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2015, pp. 1533–1542.

[236] B. Alsallakh, L. Micallef, W. Aigner, H. Hauser, S. Miksch, and P. Rodgers, "Visualizing sets and set-typed data: State-of-the-art and future challenges," in *EuroVis - STARs*, R. Borgo, R. Maciejewski, and I. Viola, Eds. The Eurographics Association, 2014.